Bayesian Networks for Health Care Support

By: Nargis Pauran

A thesis submitted for the degree of Doctor of Philosophy, 2015

Risk & Information Management (RIM) Research Group,
Department of Electronic Engineering and Computer Science,
Queen Mary University of London,
Mile End Road, London, UK, E1

Declaration

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

---------- ----------

Nargis Pauran Date

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr William Marsh for the continuous support of my PhD studies. I would like to thank to my family members for their warm encouragement as well as the generous financial support throughout my PhD studies. Finally, I gratefully acknowledge the 6 month's fund provided by ImpactQM.

Abstract

Bayesian Networks (BNs) have been considered as a potentially useful technique in the health service domain since they were invented. Many authors have presented BNs for managing health care and waiting time, predicting outcomes, improving treatment recommendation process and many more. Despite all these development effort, BNs have been rarely applied to provide support in any of these clinical areas. This thesis investigates the use of BNs for analysing clinical evidence data from observational studies, currently considered the type of study proving the weakest evidence.

It begins by investigating challenges around the analysis of data and evidence faced by health professionals in health service. It then discusses the importance of observational studies to understand how disease, treatments and other clinical factors interact with each other. Further it describes the various techniques, such as using statistical inference methods and clinical judgements, available to justify any discovered interactions. In contrary to Frequentist approaches, Bayesian Networks can combine knowledge and data to derive evidence of relationships between different factors.

This thesis proposes a novel way to combine knowledge and observational data in Bayesian Networks to derive evidence for clinical queries. Firstly, it shows how to construct and refine a Bayesian Network model by performing hypothesis tests to check which out of a number of experts' judged causal relations between a set of domain variables are plausible for the available observational data. Secondly, it proposes techniques to evaluate the strength of all plausible relations/associations. Finally, it shows how these techniques are combined into a novel data analysis method for answering clinical queries by combining knowledge with data. In order to illustrate this method this thesis uses a case study and data about the operation of a multidisciplinary team (MDT) that provided treatment recommendations to cancer patients, at Barts and the London HPB Centre over five years. In summary, the case study shows the potential for the method and allows us to propose ways to present results in a comprehensible format.

To my family

Table of Contents

Glossary of Abbreviations

BN	Bayesian Network
MDT	Multidisciplinary Team
EBM	Evidence-Based Medicine
RCT	Randomised Controlled Trial
SORT	Strength of Recommendation Taxonomy
NICE	National Institute for Health and Clinical Excellence
CI	Confidence Intervals
BF	Bayes Factor
MCID	Minimum Clinically Important Difference
DAG	Directed Acyclic Graph
CPD	Conditional Probability Distributions
CPT	Conditional Probability Table
JPD	Joint Probability Distribution
CI	Conditional Independence
NPT	Node Probability Table
MLE	Maximum Likelihood Estimation
EM	Expectation-Maximisation
BIC	Bayesian Information Criterion
AIC	Akaike Information Criterion
GS	Greedy Search
NPC	Necessary Path Condition
PBN	Prognostic Bayesian Network
MCMC	Markov Chain Monte Carlo
MDL	Minimum Description Length
BDR	Bronchodilator Response
LOS	Length of Stay
PBNs	Prognostic Bayesian Networks
MDL	Minimum Description Length
BDR	Bronchodilator Response

TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
PPV	Positive Predictive Value
NPV	Negative Predictive Value
CC	Correlation Coefficient
ROC	Receiver Operating Characteristic
AUROC	Area Under the Receiver Operating Characteristic Curve
HL	Hosmer-Lemeshow

List of Figures

List of Tables

Chapter 1

Introduction

Clinicians and other health care professionals use research studies to understand the domain. The design of a study is considered to be crucial for determining the strength of the evidence resulting of the study. Experimental studies such as randomised controlled trials (RCTs) use randomisation to decrease the effect of confounding, and provide the highest ranked evidence. Due to this most attention has been given to experimental trials rather than observational studies. However, experimental trials are not suitable for all questions of interest, and in addition, are sometime impossible to conduct since the time and cost require are often very high. Thus, an important research challenge within the health service domain for today is to produce strong evidence from observational studies.

Every study uses a technique for assessing the strength of the results that it generates. For observational studies, two common techniques are: using P values, and Confidence Intervals. Both these statistics state if the result derived from the hypothesis tests it is statistically significant. Some judgments are then made by experts to determine how likely a change can be made on the basis of this evidence. This process can become difficult since studies reporting these measures rarely indicate how they should assist in managing complex issues.

The analysis of observational data requires the use of a model, such as a multivariate regression. Bayesian networks (BNs) are well known as expert systems but can also be used to model data. A BN is a probabilistic model that represents the probabilistic relationships and conditional dependencies among variables. A BN allows probabilistic inference to be performed coherently, using the law of probability. Also a BN has the

capability to represent associations elicited from experts as well as from data, and this makes it perfectly suitable for causal modelling.

Many BNs have been developed to provide support to various clinical tasks including diagnosis, treatment selection, risk analysis and health care management [3][4][5][6][7][8]. However, their application to regular practice is still rare. Numerous studies have mentioned the existing methods for justifying the use of BNs appeared to have drawbacks [9][10]. Specifically, these methods are mostly for showing how accurate a BN model is in predicting the states of one outcome whereas in most health care domain the existing framework of clinical evidence from observed data is for providing decision support to queries regarding multiple outcomes. This thesis proposes a novel way to use a BN model to address clinical queries. . The initial interest here is to form a BN model for representing causal relations by combining the knowledge of experts and data found from an observational study.

Unlike most approaches, there is no need to construct a full model, instead the relations for each variable can be considered in turn to establish if the knowledge based causal relation show the plausibility of existence for the available data. Secondly, this thesis proposes to use the data of plausible relations of the BN model for assessing the strength of each of these relations. Further, it demonstrates that Bayesian analyses on findings from this assessment generate evidence, allowing more confident support for queries by health professionals.

The method is introduced using a case study and data collected from meetings of a Multidisciplinary Team (MDT) meeting process that treats patients suffering with cancer or suspected to have cancer. Data about the MDT meeting process were collected from the Barts and the London HPB (HepatoPancreaticoBiliary) centre following some changes to the MDT process. By evaluating the strength of each of the associations, we examine whether the MDT process has improved treatment recommendations for these patients.

.

1.1 Hypothesis

The thesis is about using a BN to combine the use of both knowledge and data to answers clinical queries. In particular, the thesis addresses the five research hypotheses listed in Table 1-1.

Table 1-1 Research hypotheses verified in this thesis

Hypothesis	**Chapter of the thesis**
It is important and possible to answer clinical queries from observational data. It is important to propose a new method for answering clinical queries that derive from observational studies.	2, 3, 4
Data, if available, can demonstrate existence of associations in an expert constructed causal Bayesian Network model	6
Using both the knowledge of experts and data from an observational study we can form a BN to represent associations between its variables.	5, 6
For the BN model we can assess the strength of each association. The results from this assessment can then help to address a relevant query with confidence.	7
The above techniques have the potential for successfully analysing observational data found outside the clinical domain	8

1.2 Structure of this thesis

Chapter 2 discusses the potential benefits of Bayesian methods for introducing new changes in health service. We review the existing approaches to examine the effectiveness of complex health care initiatives and discuss the pitfalls of these approaches.

Chapter 3 introduces BNs and reviews existing methods for their construction, including both expert judgement and learning from data. The importance of dynamic

discretisation which extends the use of BNs to continuous variables is described. Then, the chapter highlights existing techniques to complete a BN's construction process.

Chapter 4 reports on the existing applications of BNs to medical and health care domain. The chapter surveys the types of application in the health care domain and the associated techniques of evaluation and their limitations. Overall, this survey shows that the existing techniques for using BNs in this domain are not sufficient for our aim: the analysis of non-experimental data.

Chapter 5 introduces the case study. The chapter identifies the factors relevant for the problem domain, including the assumptions that a health professional might wish to confirm. Then, drawing on existing work, it presents an initial BN for the domain that was constructed by consulting with an expert. This BN contains causal relations and is the starting point of our analysis.

The next three chapters present the main contribution of the thesis.

Chapter 6, working with the case study shows how the expert-judged causal relations in the initial BN can be assessed against data, and the results can be used to ensure that the structure of the BN model given in Chapter 5 represents only those associations that both experts and data have confirmed.

Chapter 7, again using the case study, shows how to assess the strength of each expert and data based association of the BN model and to use the results to address queries. The chapter shows how to construct the posterior distribution over the parameters of each association from data, using an auxiliary multinomial BN model. Further, it shows how this model can be used to answer questions about the modelled domain, using a Bayesian approach to uncertainty and confidence.

Chapter 8 presents the complete methodology for the analysis of data, in a way that could be applied to other studies. The chapter shows that the approach is not restricted to the domain of the case study used but can be applied to successfully model associations in any domain.

Chapter 9 provides concluding remarks of this thesis.

Chapter 2

Clinical Evidence in Health Services

Introducing changes in health service is a challenging task. It takes time, collaborative effort and energy. Before implementation it is essential to ensure that this change is in the best interest of patients, will improve the quality of care, clinically-led and based on the best available clinical evidence. Many health professionals now seek knowledge from scientific research to make their decisions on organisational design. Marston et al. categorised these studies into six study types: descriptive, taxonomic, analytics, interpretive, explanatory, and evaluative [11].

In evidence-based approach, the best way to assess the effect of an intervention is to perform a randomized controlled trial [12]. When the change in health service concerns introducing a complex intervention such a multidisciplinary team which varies in composition, frequency, and processes, this is difficult to conduct [13]. In [12] Boxer et al compared the intervention group with a non intervention group to evaluate the performance of such an intervention, and acknowledged that the approach generates bias.

Identifying factors which impact on intervention effectiveness and understanding possible disadvantages of the intervention is also not simple from research studies. Altman [14] stated that a large proportion of published medical research lacking either relevance or sufficient methodological rigour to be reliable enough to answer clinical questions. Classical measures showing associations between factors such as P-values and Confidence Interval (CI) values are hard to interpret [15][16][17]. In contrast, Bayesian inference is advantageous since it allows expertise, or prior evidence, to be integrated with evidence from data.

In this study we examine the potential benefits of Bayesian methods for introducing new changes in health service. We review the existing approaches to examine the effectiveness of complex health care initiatives and discuss the pitfalls of these approaches.

2.1 Introducing organisational changes in health services

In health services managers and health professionals regularly take challenging initiatives to make useful organisational changes. For patients suffering with complex illness such as cancer and mental illness, many have focused on replacing existing recommendation process with Multidisciplinary Team (MDT) Meetings intervention.

What is an MDT?

MDT is the short word for 'multidisciplinary team'. The key idea of an MDT is that different hospital specialists work together to provide care for a patient. In each MDT meeting the team discuss patients and make decisions about the next stage of their care [18].

A patient journey in the MDT arm

For a complex disease like cancer GP refers the patient to a clinic/hospital. During the clinic visit specialists examine symptoms, test results and other relevant factors to understand the current state of the disease. After diagnosis, the patient is discussed at an MDT meeting where doctors, specialist nurses, the oncologists, surgeons, radiologists and pathologists meet to discuss the specific case, and consider the scans, general health of the patient, the type of the cancer and the wishes of the patient to decide an appropriate course of treatment.

What is involved in the non-MDT process?

In a non-multidisciplinary setting a GP can provide health care to patients. Figure 2-1 provides a simplified overview of the pathways followed in general practice to provide effective treatments to patients.

In brief, after seeing a patient presenting with signs and symptoms, a GP may follow (a) the history of the patient, (b) any presented complain, (c) the outcomes of test(s), or (d) all listed options for diagnosis. The GP applies the diagnostic label to justify a decision to prescribe a treatment option. If no treatment is needed then the GP may discharge the patient or sent him back home to monitor. For a complex diagnostic label, the GP may decide to refer the case to a specialist in a hospital. This referral often initiates the MDT process for recommendations.

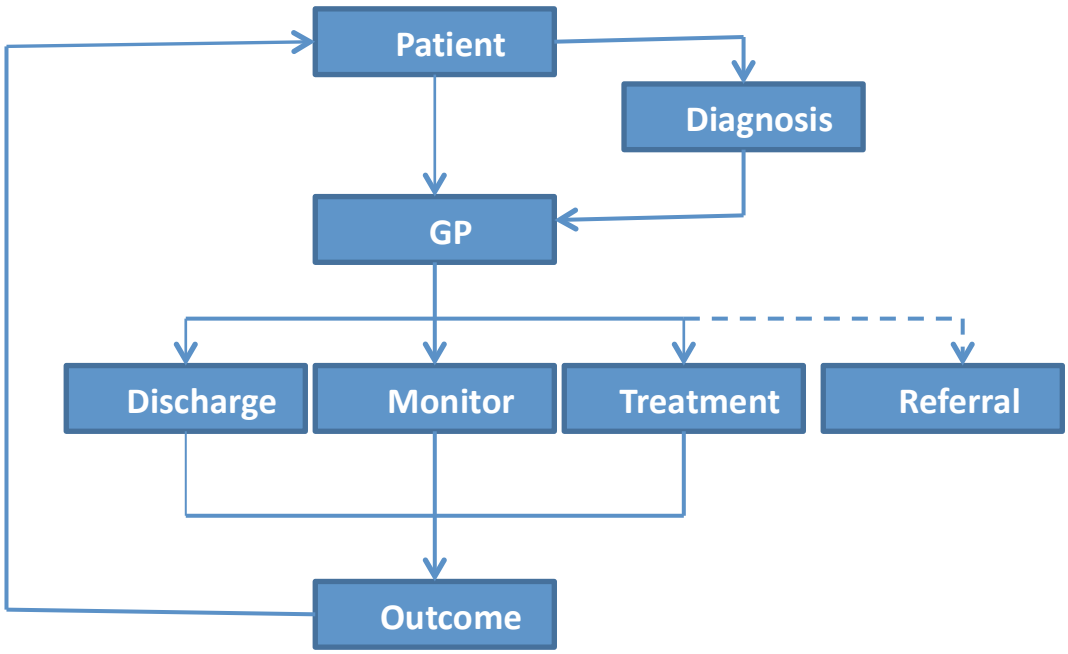


Figure 2-1 A non MDT process

Evaluating the impact of MDT

As MDT meetings require substantial administrative, human and technical resources to run successfully, it is important for the meetings to run well[19]. The following questions can therefore be used for evaluating MDT meetings:

- How long does it take to discuss each patient?
- At how many meetings is each patient discussed?

- Is the decision correct?

In [12] Boxer et al compared MDT and non MDT groups patients to understand the pattern of care for those suffering with lung cancer and reported decreased diagnosis to surgery time, more recommendations of radiotherapy (66% versus 33%, P <0.001), chemotherapy (46% versus 29%, *P < 0.001*), and palliative care (66% versus 53%, *P < 0.001*). In [20] authors reported better treatment recommendations in the MDT group (76.81% patients underwent neoadjuvant chemotherapy instead of direct surgery). In [21] the use of a multidisciplinary breast cancer evaluation programme provided important second opinions for many and led to a change in treatment recommendation for 43% (32 of 75) of the patients. In [22] authors compared the treatment of patients with inoperable NSCLC before and after the introduction of MDT, and found that the introduction of an MDT was associated with an increase in the proportion of patients being staged and a change in treatment; more patients received chemotherapy and fewer received palliative care only.

In addition to the benefits in treatment recommendations, MDT discussion can facilitate faster initial of treatment [23], reduce hospitalisation time and costs [24][25], and improve coordination of care[26]. Many authors have shown that patients who received multidisciplinary care were significantly more satisfied than others [27][28]. Also in most situations patients prefer to visit once to a clinic and receive care provided by a team of specialists then a multiple-visit approach [23].

A number of the above studies are based on before and after design [29]. Since care for cancer patients is improving over time, it may be possible that patents for recent times were staged more accurately than those diagnosed before [12]. Some MDT related research studies have shown the benefits of MDT meetings by conducting multiple concurrent organisational changes such as centralising the process [30], increasing caseload [31], and appointing new specialists [32]. Despite these adjustments the evidence for implementing an MDT process is compelling. The question of ones interest is now how to implement this process change and this is what addressed in the empirical part of the thesis.

2.2Causality

The term causality is about investigating that a cause is something that produces an effect. For the time-series data in Economics, Grenge [33] formulated causality as this - a ‘cause’ ought to improve our ability to predict an effect in a probabilistic system. In this discipline the use of structural equation models has been seen as the dominant approach to making inference about causal effects. An example of causality from Medicine is that the establishment of specific bacteria is the cause of specific infectious diseases. In Statistics a probabilistic cause is one that increases or reduces the chance that the effect will occur, and a probabilistic statement about a cause and effect gives quantitative information about an estimate of the strength and nature of that relation. It also provides quantitative information on potential effect modification and about any relation that may exist between the cause and its effect [34].

Experimental studies, such as Randomised Controlled Trials (RCTs) often provide the most trustworthy methods for establishing causal relations from data. In an experimental study (a) the analysis of causation begins with studying the effects of causes, (b) effects of causes are always relative to other causes – it takes two causes to an effect and (c) not everything is taken as a cause so focus is always on some specific fields. Such studies, while potentially highly informative, may be:

- Impossible, impractical or unethical. For example, to assess the effectiveness of photodynamic therapy (PDT) it is impossible for the clinicians to allocate the patients with brain tumours into a PDT group and a placebo group [35].

- Unnecessary if the effect of a treatment is very high. For example, the use of Imatinib (Glivec) to treat patients with Chronic Myeloid Leukaemia [36].

- Inadequate. In an RCT, the patients who participate must meet the chosen study criteria and the follow-up period is short, whereas the treatment evaluated takes place in clinical practice – where a large quantity of patients suffering from many different conditions are of concern – for a longer period

of time. Thus, the estimate of an effect of a treatment which was found in the RCT may not be externally valid or generalisable outside the context of the trial [37] [38].

- Increases the type-1errors – false positive errors – due to interim analyses. During an ongoing experiment, an interim analysis of data is held to let the investigators know whether the intervention is efficient or harmful in relation to the current placebo in order for them to decide if the study should be stopped earlier than planned. In any situation, an interim analysis that is held but hasn't been planned in advance increases the type-1 error [39].

Causality to nonrandomised observational studies has also been investigated extensively. Observational studies are based on observed data, and these data are more readily available than experimental data. As observational data become increasingly available, opportunities increases for using them. Besides, observed data describe situations that happened in the past and there are no needs for doing experiments. These make applications of observational studies higher where generalisation is an issue during discovery of causal relations or simply statistical associations.

2.2.1 Causal DAG

Directed acyclic graph (DAG) models are well-used tools for capturing causal relationships and for guiding attempts to discover these relations from data. They supply a means of extracting causal conclusions from probabilistic conditional independence properties inferred from purely observational data.

A DAG consists of nodes that represent variables, and arrows that join these variables. Given a joint distribution over an ordered set of random variables $(V_i, \dots, V_N)$, one can construct an associated DAG with the (V_i) as vertices by adding arrows to $V_{i+1}(i = 0, .., N-1)$ from the smallest subset, S_i, of all earlier variables, $V^i = (V_1, \dots, V_i)$, such that

$$V_{i+1} \coprod V^i \mid S_i$$

According to Hernan and Robins [40], a causal DAG is a DAG in which:

i. the lack of an arrow from V_1 to V_2 can be interpreted as the absence of a direct causal effect of V_1 on V_2 (relative to other variables on the graph)

ii. the inclusion of the measured variables implies that the causal DAG must also include unmeasured common causes.

One description of the idea of 'cause' relates it to an intervention or forced changes. If A causes B, then forcing A to a new value causes B to change. This is relevant to decision making when there is a need to change B but it cannot be done directly. Pearl [41]and other have popularised causal modelling and have developed a principled way to predict the outcome of interventions. These models assume that the direction of causal relations is known from some source other than data.

Also, with two variables A, B, it is not possible to tell from data (i.e. statistics) whether A causes B or B causes C or neither. The third cases arises from a possible unknown third variable U, which causes both A and B and therefore give rise to a correlation between A and B.

2.3 Frequentist inference and interpretation

In the frequentist method, an investigator performs a statistical test – e.g. a hypothesis test – on a sample of a target population and uses the results of the test to make inference about the population. The method mostly utilizes *P-values* and confidence intervals (CIs) to measure the results of the test. Because these measures are easy to misinterpret they can easily lead to an invalid conclusion when a clinician considers them to decide on clinical application [15]. I discuss about the use of the measures of frequentist more in Sections 2.3.1 and 2.3.2. In contrast to the frequentist method, the Bayesian methods of statistical inference are rather simple and highly applicable

alternative for assessing the results of statistical tests [42]. This method is described in Section 2.4.

2.3.1 Null hypothesis, P-values and Confidence intervals

The frequentist method uses hypothesis tests and considers the null hypothesis. For a hypothesis such as: *is chemotherapy with surgery more effective as a treatment for the patients with stomach cancer than the chemotherapy alone*, the null hypothesis,H_0, particularly states: *there is no difference in effect between the treatment options*. An investigator tests a null hypothesis by calculating the P-value.

There are many definitions available for a P-value.These definitions are placed in relation to either the test of significance of Ronald A. Fisher or the hypothesis testing method of Jerzy Neyman, and Egon S. Pearson. Here we have given two definitions to demonstrate the difference:

> "The P-value is the probability of observing data as extreme as, or more extreme than, the data actually observed assuming that the null hypothesis is true". [43]

> "A p-value is the probability of finding a result as extreme, or fantastic, or disappointing, as the one return by a statistical test". [44]

A P-value of 0.05 is the cut-off for rejecting or not rejecting the null hypothesis [45][46]. If the measured P-value of the test is below the cut-off it indicates that the result is statistically significant [47] – that is, considering the hypothesis above, any difference in effect between the treatment options is likely to be real and not to have occurred by chance [48]. Conversely, a P-value which is equal or greater than 0.05 informs the investigator that the evidence is not sufficient to reject the null hypothesis so the result of the test is not statistically significant and any difference may be due to chance.

However, a P-value less than the cut-off point is not the evidence to support the hypothesis. According to Akobeng [45], a *"p-value of <0.05 should not be regarded as*

'proof' that an intervention is effective, and a p-value of ≥0.05 does not also mean that the intervention is not effective". Further, the P-value does not give information about whether the result of the test is suitable to be placed into clinical practice. Many have therefore suggested that a study should use confidence intervals (CIs) to determine the clinical importance of the result [42]. There are now also requirement that to submit a clinical paper the authors must include both CIs with the P-values in their results [49][50].

A CI measures the significance of a result and the magnitude of effect of the intervention that is under consideration [42]. Instead of defining a probability, the CI estimates a range of plausible values within which the 'true' value is expected to be observed, and the use of the 95% CI, the measure that is most common in the studies, indicates that 95 out of 100 times the true value lies within the range [15]. The width of the CI can indicate (1) the precision of an estimate – the narrow the CI the better is the precision – and, (2) the amount of error in an estimate.

As a method of reporting the statistical significance of results, CIs are known to be easier than a P-value.[45]. In brief, along with the information about the statistical significance, a CI helps an investigator to determine when an effect is clinically important [48] [51][52].

2.3.2 Criticisms of the use of P-values and CIs

The use of both P-values and CIs have been criticised for their improper use and misinterpretation with regards to the results of clinical research studies.

In [53], Steven has stated twelve misconceptions that exist among investigators when they decide to interpret a P-value that arises from a two-group randomized controlled trial. Some of these misconceptions also exist among the investigators when they interpret the P-values that arise from an observational research study. Fenton and Neil [54] also present a summary of the problems with the use of P-values, drawing on the book *The Cult of Statistical Significance* [55]. This thesis includes some of the criticisms which the authors of the above studies have mentioned, and in addition,

included few more that other researchers have made during their discussion regarding the use of the P-values and CIs.

Specifically, the following criticisms are found regarding the use of P-values:

- Provides misinterpreted assessment of the null hypothesis.

If a P-value is 0.05, an investigator often misinterprets the result stating that there is a 5% probability that the null hypothesis or no relation is true [56][57]. However, R.A Fisher has introduced the P-value as a rough numerical guide which one can follow to evaluate the strength of evidence against the null hypothesis. His proposal in relation to a P-value less than 0.05 was to suggest an investigator to repeat the test and make conclusions based on the results of subsequent tests. Besides, it is not possible to equate the P-value with the probability of the null hypothesis since an investigator only calculates the P-value considering that the null hypothesis is true [56].

- Makes one to focus less on the problem of interests.

A smaller P-value is only to show more evidence against a null hypothesis. But from an investigator's point of view the exclusive focus on the null hypothesis is not always interesting [54]. Goodman [58] demonstrates that deriving the strongest evidence sometimes requires a different scope, addressing a question that needs a composite hypothesis in order to answer it properly.

- Generates confusion because the use of cut-off level is arbitrary.

A P value of < 0.05 is commonly taken as statistically significant, but the use of this threshold for the P-value has been considered to be arbitrary [46][59]. For an investigator it is difficult to propose an interpretation of a P-value that is near to 0.05; it turns out that during an interpretation any previous evidence or the judgment influences the suggestion. For example, if a P-value is 0.06, the value is a short distance from the significant level and therefore, the investigator can suggest the interpretation that the P-

value is showing is that the result is almost statistically significant or he can simply say that there is no evidence of a relation [42].

- Accelerates inconclusive assessment of clinical relevance.

The P-value does not provide the information that is more important to a clinician, namely, the clinical significance or relevance of a relation [54]. What clinicians are interested in is to know about the magnitude of an effect and therefore, they can intuitively misinterpret statistical significance as clinical significance.

- Provides wrong interpretation of the used data.

"The P-value is not the probability of the observed data under the null (chance) hypothesis, because the P-value includes the probability of more extreme data". [56]

- Provides misleading information for the tests of significance.

Two tests based on the same observed data do not always give us the same P-value [58]. Goodman [56] has further demonstrated that if two trials are run, and the sizes of the trials are the opposite – one is large whereas the other is small, even if the P-values for both are 0.05 the evidence against the null hypothesis in these cases will be different. This tells us that an investigator's assumption that the identical data means the identical evidence will not always be acceptable.

- Overemphasised on the use of the cut-off level.

When interpreting the P-value the investigator's emphasis on the 0.05 threshold value is regarded more strongly than it should be [60]. Since a P-value of 0.05 only corresponds to the minimum Bayes Factor (we will discuss more about Bayes Factor in Section 2.3) of 0.15, it represents at best moderate evidence against the null hypothesis. Sterne et al. [46] have mentioned that a strong evidence against the null hypothesis comes from a P-value which is much smaller than 0.05.

- Inappropriate for an inductive inference.

The P-values represent deductive inference [42]; a hypothesis is initially held and tests are then performed to check if the observations are consistent with the hypothesis [61]. In contrast, strong association inference in clinical practice requires inductive inference where the clinicians first make observations and then decides which hypothesis is likely for the observations [62]. Since their interest is in knowing the probabilities of effects based on the observed data, the clinicians often incorrectly interpret P-values using inductive inference [61][63].

- Suitable only to support dichotomous outcome.

The emphasis of a P-value is on the strength of evidence considering that only dichotomous 'reject' or 'fail to reject' outcomes.

The criticisms made regarding the use of CIs are as follows:

- For the 95% CI, an investigator often interprets this to imply that there is a 95% probability that the true relation lies with the 95% CI [64].Whereas what 95% actually means is that when the same test is repeated many times and the CI is calculated for each, then 95% of such intervals will include the true relation [64], and many clinicians ignore this distinction during their interpretations.

- While making an interpretation, clinicians do not always consider the implications – the importance for applying into clinical practices – of the range of values in the interval [46]. They prefer to use CIs to examine significance and the 95% CI usually uses $P < 0.05$. A relation is classified as significant when the 95% CI excludes the null hypothesis of no relation [15]. Regarding this drawback with CIs, Goodman [61] states "their impact on the interpretation of research is unclear".

- Clinicians do not always prefer to make healthcare decisions with 95% confidence. Their interest in confidence can vary in relation to the effect of an

association between the factors. For example, a clinician may be interested in knowing whether an intervention has an 85% probability of showing an effect.

- The use of a CI can lead to a misleading conclusion. With the reference to the study [65] Shakespeare et al. [15] explained that the results given suggesting the effect is not being statistically significant, do not rule out a potential benefit completely and therefore that, a conclusion of 'no effect' is misleading.

- The relationship between the width of a CI and the sample size of the test is not always linear; in general, investigators need to increase the sample sizes by a factor of four to halve the width of the CIs [51]. Therefore, if a clinician intuitively thinks that a linear relation exists, erroneous decisions can result.

2.4 Bayesian inference

Bayesian inference method can overcome many limitations of the frequentist and offer many advantages when it comes to evaluating the strength of evidence. In particular, these advantages can be summarised by the following points:

- Bayesian inference calculates the evidence in favour of a null hypothesis [66][60][67].

The usual interpretation of significance tests can be used only to reject hypotheses and do not offer an assessment of the strength of evidence in favour of the null hypothesis. Bayesian methods of statistical inference let us calculate what is really required as evidence – including the cumulative impact of different evidence – and this is not $P\ (D|H)$ but $P\ (H|D)$ by using Bayes Theorem:

$$P(H|D) \propto P(D|H) * P(H) \qquad 2.1$$

Here, $P\ (H)$ specifies the prior probability regarding the uncertainty for the hypothesis and $P(\ D|H)$ is the likelihood of data which specifies the probability of the data given the hypothesis. These two then combine in Bayes' Theorem to give $P(\ H|D)$ which is the posterior probability of the hypothesis given the observation that is made from the data.

- It can handle uncertainty in moeling and draw inference about a relation.

While investigating the effect of one variable on another one needs to think about several other possible covariates. It is not often clear if the chosen set is the right set of covariates, and this confusion generates uncertainty that should be taken into account during this investigation. Similarly, other functional and distributional assumptions may lead to different estimates of quantities of interest, and again, one would like to take account of uncertainty about these assumptions within the estimation process. The Bayesian inference method allows doing this in a natural way by averaging over the candidate models with their posterior probabilities as weights [66][57][68].

- By giving the probability for a hypothesis on the basis of the data Bayesian methods permit inductive inference, which are more appropriate for assessing cause-effect relations. [62][69].

- During hypothesis tests, the use of Bayesian methods enable a researcher to measure the strength of the evidence by calculating Bayes factors (BFs) [56][61][70].

A Bayes Factor (BF) is the ratio of the probabilities of two competing hypotheses. For example, if we have two hypotheses: H_1 and H_2, the BF computes by taking the ratio of the conditional probability of the hypothesis, H_1, given the data D and the conditional probability of the competing hypothesis, H_2, given the same data D , from the equation as follows:

$$BF = \frac{P(H_1|\mathrm{D})}{P(H_2|\mathrm{D})} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)} \qquad 2.2$$

When the priors for the hypotheses – $P(H_1)$ and $P(H_2)$ – are equal, the Bayes factor of two competing hypotheses is

$$BF = \frac{P(D|H_1)}{P(D|H_2)} \qquad 2.3$$

In words,

Posterior odds = Bayes factor × prior odds

The use of Bayes factor can quantify the evidence for one hypothesis relative to another [71] and can suggest which hypothesis is better for predicting the data that an investigator observes. According to Goodman [58], the use of Bayes factors lets the investigator make a clearer distinction between evidence and error, and enables him to derive a proper measure for the evidence. The magnitude of the Bayes factor calculated from the observed data also informs us about the strength of the evidence [56]. For this purpose, Jeffreys [72] has recommended a scale (Table 2-1); an interpretation of the Bayes factor based on the scale helps to decide whether the data we have to hand support one hypothesis over another and in what degree.

Table 2-1 Jefferys' scale of evidence for Bayes factors

Bayes factor	Strength of Evidence
> 100	Decisive evidence for H_1
30 - 100	Very Strong evidence for H_1
10 - 30	Strong evidence for H_1
3 - 10	Substantial evidence for H_1
1 - 3	Anecdotal evidence for H_1

According to the scale, if the Bayes factor for H_1 against H_2 is 1.5, and then the interpretation indicates that the observed data are 1.5 times likely to have occurred under the hypothesis H_1 than under the hypothesis H_2 . This BF corresponds to the category: Anecdotal evidence for H_1, which specifies that although the BF favours the hypothesis H_1, we do not have strong evidence from the observed data to reject or accept either hypothesis.

The use of Bayes Factor provides the flexibility – such as linking evidence, supporting the rare co-ordination – during Bayesian analysis. In particular, these advantages can be summarised by the following points:

- A large number of hypotheses is sometime needed to test a complex relationship. To determine the strength of the relations results from all these tests needed to be used to ensure a complete assessment. An investigator can take the cumulative of the BFs from the tests during an analysis.

- A large P-value can tell that the result found is not statistically significant. But one can simply find a large P-value just because there is less or no indication of that particular relation in the available data. The use of priors allows taking account of this uncertainty within the Bayes Factor.

2.5 Summary

This chapter examines the potential benefits of Bayesian methods for introducing new changes in health service. It reviews the existing approaches to examine the effectiveness of complex health care initiatives and discuss the pitfalls of these approaches.

Chapter 3

Review of Bayesian Networks

The chapter discusses the basic components of Bayesian Networks (BNs). The chapter starts with an introduction to BNs and then discuses, in Section 3.1.1, how a BN can model both discrete and continuous variables. Section 3.1.2 discusses the methods developed to perform inference in a BN. Section 3.2 provides a review of the methods used to construct a BN, with Section 3.3 covering the tools available for constructing BNs.

3.1 Introduction to Bayesian networks

The theory of BNs has been developed since the early 1980s, building on early work by Pearl [41], Jensen [73], Lauritzen and Spiegelhater [74] and others.

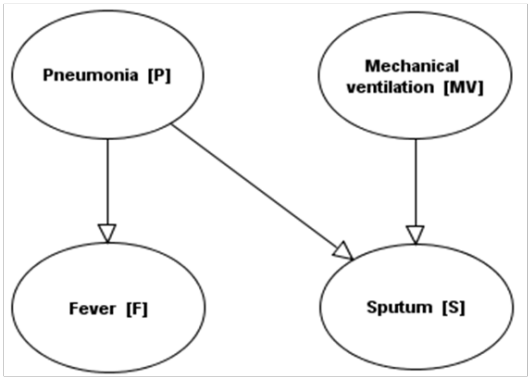
A BN, also known as a causal probabilistic network or a belief network, is a graphical model that represents probabilistic relationships among a set of variables. They consist of two parts: a graphical structure and a set of parameters. The structure is to represent a directed acyclic graph (DAG) and the parameters are to determine the Conditional Probability Distributions (CPDs) for the variables.

The structure of a BN is formed by the variables – known as nodes – and links that connect the variables. The variables represent the factors (or the propositions) that we find as relevant for the domain being modelled and the links convey information about dependency relations between the variables. Usually, the relations within the structure are expressed by using the wording of family relations: for example, to describe the

relation formed by the link from a variable *X* to another variable *Y*, we often prefer to say that *X* is a parent of *Y* or *Y* is a child of *X*.

The purpose of the Conditional Probability Distributions (CPDs) in a BN is to quantify the strength of the relationships that are defined within its structure. The Conditional Probability Distributions can be defined as: (a) Conditional Probability Table (CPT) – for discrete variables and (b) Conditional Probability Distributions (CPD) –for the continuous variables. The BN that has both discrete and continuous variables is known as a 'Hybrid Bayesian Network' [75]. The CPD of a variable has a collection of parameters, the number of parameters being determined by the type of considered variable.

Figure 3-1 is a BN for the diagnosis of pneumonia. This model is based on the discussion in [76], but represents the development of the disease in a simplified manner. Since all the links of the BN are chosen to show that the parent variables have strong influence on the child variables, this can be regarded as a Bayesian Network representing strong associations. For example, the link *Pneumonia* → *Fever* shows that *Pneumonia* is a cause for *Fever*. The BN has two parents for the variable *Sputum*; *Pneumonia* and *Mechanical ventilation* regard as two separate, but interrelated, causes for increased (High) production of sputum.



Pneumonia [P]	
Yes	0.75
No	0.25

Mechanical Ventilation [MV]	
Yes	0.65
No	0.35

Fever [F]	Pneumonia [P] = Yes	Pneumonia [P] = No
Yes	1	0.2
No	0	0.8

Sputum [S]	Pneumonia [P] = Yes		Pneumonia [P] = No	
	Mechanical Ventilation [MV] = Yes	Mechanical Ventilation [MV] = No	Mechanical Ventilation [MV] = Yes	Mechanical Ventilation [MV] = No
High	0.9	0.7	0.7	0.1
low	0.1	0.3	0.3	0.9

Figure 3-1 A BN formed of four variables

One important property of BNs is their ability to represent the Joint Probability Distribution (JPD) for all the variables in a compact form [77]. The traditional approach, i.e. the chain rule, requires a full specification of the probability distributions. In this traditional approach, the full JPD of a probabilistic model with n random variables$X_1, X_2, \ldots., X_n$ is as follows

$$P(X_1, X_2, \ldots., X_n) = \prod_{i=1}^{n} P(X_n | X_1, X_2 \ldots \ldots, X_{i-1}) \qquad 3.1$$

This approach is complex and has the potential to introduce a high number of probability entries [78]. In contrast, the framework of a BN reduces the complexity inherent in the full joint probability distribution by reducing the number of probabilities

of an inference. In particular, given the structure, the JPD for the BN is the product of all conditional probabilities specified in the BN:

$$P(X_1, X_2, \ldots., X_n) = \prod_{i=1}^{n} P(X_i \,|pa\,(X_i)) \qquad 3.2$$

where $pa\,(X_i)$ are the parents of variable X_i in the BN. The conditional probabilities are therefore determined by $O(2^{|pa(X_i)|+1})$ parameters instead of $O(2^n)$ that they would otherwise require if the JPD were computed directly using Equation 3.1. Using Equation 3.2, the BN saves (a) the space that requires for storing the parameters and (b) the time that requires for computations.

Based on the BN in Figure 3-1 the JPD by using Equation 3.1 computes as:

$$P(P, MV, F, S) = P(P).P(MV|P).P(F|P, MV).P(S|P, MV, F) \qquad 3.3$$

Considering the necessary conditional independence assumptions, the JPD by using Equation 3.2 computes as:

$$P(P, MV, F, S) = P(P).P(MV).P(F|P).P(S|P, MV) \qquad 3.4$$

This representation allows us to determine the conditional probabilities for the large CPT, i.e. the CPT for the variable *Sputum*, with 2^{2+1} i.e. 8 parameters rather than 2^4 i.e.,16 that is required otherwise.

Further, we can calculate the marginal probability distribution for *Pneumonia* from:

$$P(P) = \sum_{MV,F,S} P(P, MV, F, S) \qquad 3.5$$

$$= \sum_{MV} \sum_{F} \sum_{S} P(P)P(MV)P(F|P)P(S|P, MV)$$

$$= \sum_{F} (P(F|P) \sum_{MV} (P(MV) \sum_{S} P(S|P, MV)))$$

Before entering any evidence, based on the prior probabilities on variables, the probability that a patient is producing high sputum is 74.5%. We can calculate this using equation 3.6.

$$P = \sum_{P} \sum_{MV} P(S = High|P, MV)\, P(P)P(MV) \qquad 3.6$$

$$= 0.9 * 0.75 * 0.65 + 0.7 * 0.75 * 0.35 + 0.7 * 0.25 * 0.65 + 0.1 * 0.25 * 0.35$$

$$= 0.43875 + 0.18375 + 0.11375 + 0.00875$$

$$= 74.5\%$$

If we get the information that the patient is not receiving mechanical ventilation than this situation will reduce the probability of producing high sputum to 0.55 from Equation 3.7.

$$P(S = High|MV = No) = \sum_{P} P(S = High|P, MV = No)\, P(P) \qquad 3.7$$

$$= 0.7 * 0.75 * 0.35 + 0.1 * 0.25 * 0.35$$

$$= 0.525 + 0.025$$

$$= 0.55$$

This probability is lower than earlier one, since we are now sure that the patient did not receive any mechanical ventilation. The right side BN in Figure 3-2 demonstrates that the evidence we enter into *Mechanical ventilation* does not cause any change in *Pneumonia* since the variables are conditionally independent given *Sputum*.

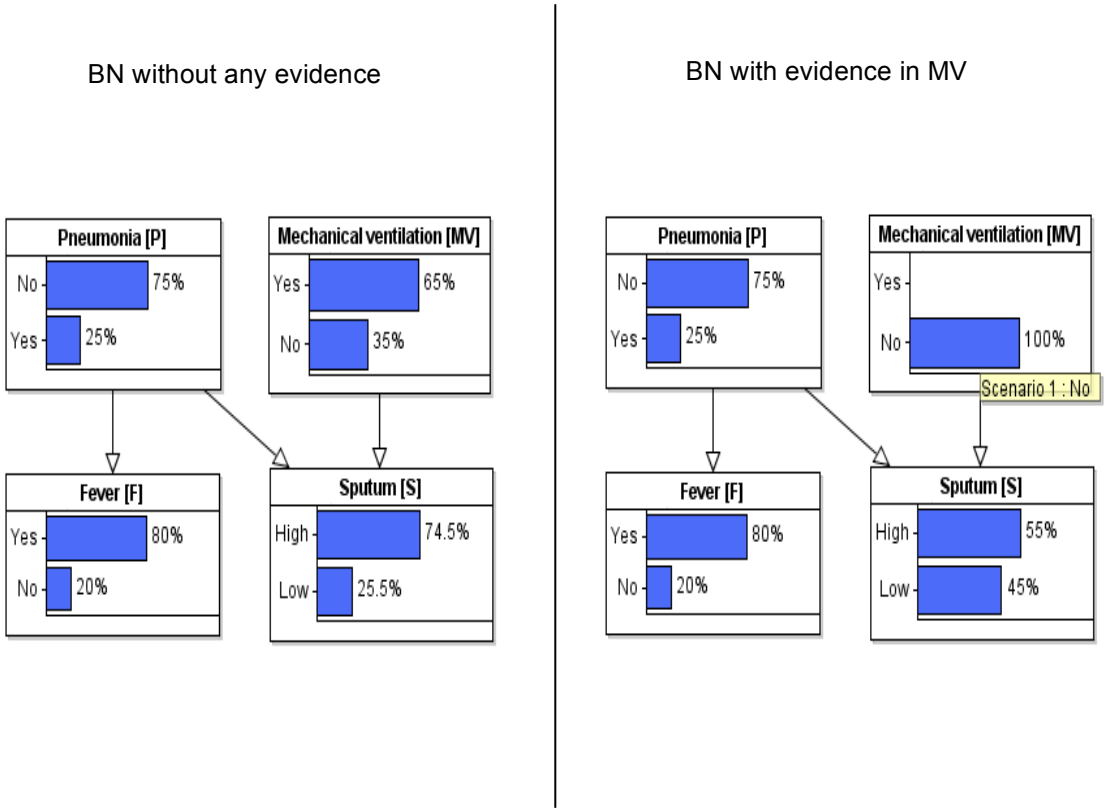


Figure 3-2 The probability of 'High' sputum for no evidence (in the left BN) and the probability of 'High' sputum for no mechanical ventilation is observed (in the right BN)

3.1.1 Static and dynamic discretisation methods

A BN can include both discrete and continuous variables [75]. A discrete variable has a finite, usually small, set of discrete values and in contrast, a continuous variable has an infinite number of values: that is, the variable can take any value between any two points on a scale. However, BNs generally require a continuous variable to be discredited in order to reduce the number of distinct values by dividing its whole range into a finite set of disjoint intervals [79].

Discretisation methods mainly fall into two types, namely: static discretisation and dynamic discretisation. In a static method, the intervals do not change: that is, remain static, while in a dynamic method the intervals change depending on the assignment of probability. Traditionally, a static discretisation is performed in two steps: 1) the user defines the number of intervals for the range of the given continuous variable and 2) the set of cut-off points is then determined in order to assign the probabilities. There are a number of methods that can help users to make the best choice when defining the number of intervals [80][81]. In [81], Clemen has explained 'the bracket median method' where the values of a continuous variable are divided into n equal- width intervals and the median of each of the intervals is used to label the interval. In each of the n intervals the assigned probability is $1/n$.

The number of intervals used influences the accuracy of a BN [82] [83] and also for the level of complexity for a computation [84] – resulting in a longer time to calculate the conditional probabilities in the BN as the number of intervals increases. As a result, to limit the complexity, a user likes to define a minimum number of intervals; this leads to more intervals in the areas where the probability is expected to be varying fastest and fewer, wider intervals where the probability is expected to be more nearly constant. Since in a static method the user would have to understand in advance what the probabilities will be in each area of the range, this leads to more intervals than are essentially required [82].

In [84], Kozlov and Koller attempted to overcome the limitations of static discretisation methods and described a dynamic discretisation method instead. The work shows that an iterative algorithm that can be used to perform discretisation, varying the intervals depending on the evidence observed. Another dynamic discretisation method was later proposed by Neil and his colleagues [79], and the general outline of the method can be found in [83] [79]. In brief, the discretisation starts by considering the whole range of the variable and then recursively divides the range into two intervals until an acceptable level of accuracy is obtained. In each of the iterations the conditional probabilities are calculated and inference is performed to update the probabilities given the observed evidence. Although the method of [79] is influenced by the work of Kozlov and Koller, but the algorithm is simpler and yet capable of achieving higher accuracy.

Figure 3-3 demonstrates dynamic discretisation in a continuous node with the distribution *Normal (mean=5, variance =1)* and the distribution *Normal (mean=10, variance=5)* with green and blue colour. The node has been discretised following Neil's method. The result clearly shows wider intervals in those areas where the probability is more constant – at the centre and extremes of the normal – and narrow intervals elsewhere.

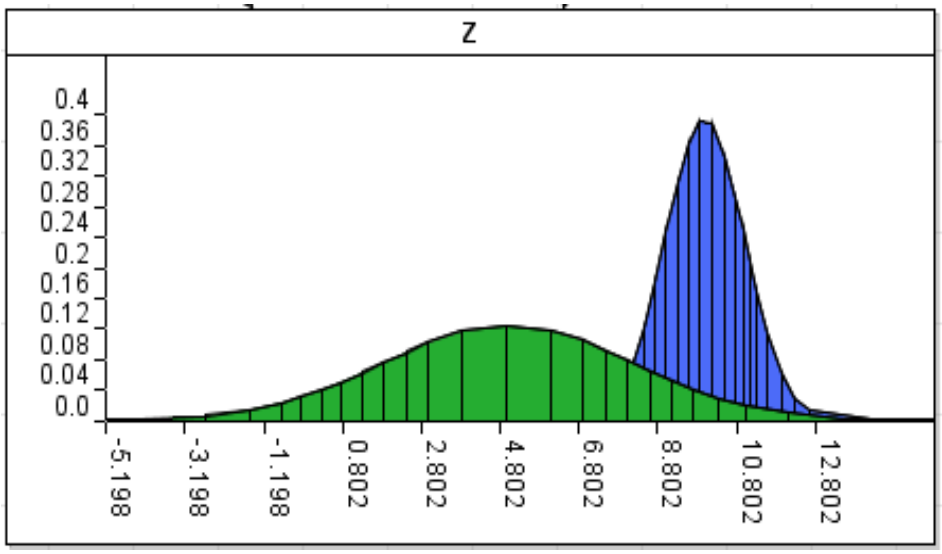


Figure 3-3 Dynamic discretisation in a continuous node with Normal (5, 1) and Normal (10, 5)

In general, a dynamic discretisation method let us to model a BN by producing more accurate intervals in the areas that matter the most. In addition to this, the adjustments to the intervals in response to new evidence also help to ensure that a greater accuracy is obtained in the BN. Because of these advantages dynamic methods are most suitable for incorporating continuous variables with those that have discrete states [85], and consequently, have been applied by many to deal with clinical problems where some factors of interests have to be measured using a continuous scale. In this thesis, the BN needs both discrete and continuous variables, so we have preferred to use a dynamic method for discretisation.

3.1.2 Inference in Bayesian networks

A Bayesian Network can update the probability of any unknown variable. In particular, given evidence into some variables the model calculates the posterior probabilities for the variable of interest. However, many BNs are complex, which makes it impossible to perform Bayesian inference calculations manually. Due to this, an increasing number of researchers have turned their attention to develop efficient methods of Bayesian inference. Some of the classic methods for exact inference are summarised in Table 3-1.

Table 3-1 Methods of Bayesian net inference

Author	Year	Description
Perl [86]	1986	A message propagation algorithms whereby the probability distributions for each variable update in response to observations of one or more variables. Pearl's original algorithm applies only to networks that are trees but similar algorithms have since been used for approximate inference in general BNs.
Shachter [87]	1988	The algorithm reverses links until the explanation of the probabilistic query is obtained from the network
Lauritzen and Spiegelhater [74]	1988	The 'Junction tree' algorithm initially transforms the BN into a tree, clustering some variable together. Inference is then done by message passing, using an algorithm similar to Pearl's.
Zang and Poole [88]	1994	A variable elimination algorithm whereby variables eliminate after summation.

3.2 Construction methods for Bayesian nets

BNs provide an ideal mechanism to model problems that involved uncertainty. At the same time, a network can be formulated to combine all the relevant domain information in an intuitive way. The links can represent cause-effect relations between variables so that it is possible to design the relations in a BN from the understanding of domain experts.

The steps necessary to construct a BN model successfully are:

- Step 1: Identify the important variables to be included in the model.
- Step 2: Define the relationships between variables to complete the qualitative part of the model.
- Step 3: Assess the prior and conditional probabilities to create the quantitative part of the model.

Identification of variables

Usually the expert finds it easy to determine the relevant variables. An extensive analysis of the purpose of the network under construction also helps to find them [89]. In [90] Heckerman suggested the following four checks for selecting variables:

- Correctly identify the goals of modelling (e.g. prediction versus explanation versus exploration);
- Identify many possible observations that may be relevant to the problem;
- Determine what subset of those observations is worthwhile to model ,and
- Organise the observations into variables that contain mutually exclusive and collectively exhaustive states"[90].

The values (also known as states) of any variable (discrete or continuous) need to represent the level of details that is required for the proposed use of the BN. In a discrete variable, the values must satisfy the 1) mutually exclusive and 2) collectively exhaustive properties. The mutually exclusive property ensures that in the variable no one value overlaps with another while the collectively exhaustive property ensures that all the possibilities for the variable are considered within the values of the variable.

Finally, a BN is usually modelled based on the following three categories of variables [73]:

- Hypothesis variables: these variables' values are either not observable at all or can only be observed at an unacceptable cost. The primary task in modelling a BN is to identify these variables.

- Information variables: these variables' values are possible to observe and have relations with the hypothesis variables.

- Mediating variables: these variables are included to meet special purposes – such as to simplify the conditional probabilities in the BN.

Defining the relationships

In addition to causal relations, one can also capture various other relation types – such as deterministic, statistical and analogical [91]–by drawing links between the variables for the qualitative part of the BN. In [91], Neil et al. suggested that the direction of a link should be defined carefully and recommended five types to reasoning, referred as idioms, to consider different problem descriptions while constructing the BN. These are: definitional/synthesis, cause–consequence, measurement, induction and reconciliation.

- Definitional/Synthesis idiom – this idiom can be used to represent relations that are deterministic or definitional such as $X = Y \times Z$. Here the synthetic node X is determined by the values of the parents Y and Z. Thus, the use of synthetic nodes can ease calculation or understanding, and create hierarchies of sub-attributes to define complex super attributes.

- Cause-consequence idiom – this idiom can be used to model a causal process in terms of the relationships between its causes and consequences. The relationships are represented using arrows, where the direction of an arrow indicates causal direction, given by knowledge that change that occurs in the cause represented by the causal variable affects the phenomenon modelled by the consequence variable. The conditional probability table within the idiom

models the uncertain dependencies between causes and consequences.

- Measurement idiom – this idiom is used to model the limits of our ability to make measurements accurately. The steps involved within the idiom are (1) the true value that is to be estimated by the measurement is distinguished from the measurements, and (2) the way the measurement instruments interact with the measured entity is modelled. The use of such an idiom helps to explain away false positive results.

- Induction idiom – this idiom represents the process of performing inductive reasoning, where a set of similar entities is used to obtain an estimate about a future entity.

- Reconciliation idiom – this idiom reconciles independent sources of evidence about a single attribute of a single entity, where these sources of evidence have been produced by different methods. In addition, it combines uncertain definition model with causal inference model and combines information from various causal models.

Assessing the probabilities

The final step of the construction process is to assess the prior or conditional probabilities for each of the variables. The conditional probability, $P\left(X\middle|pa(X)\right)$ defines the probably distribution over the states of each variable X given its parents $pa(X)$. If the variables are discrete, the distributions represent as Node Probability Tables (NPTs) and otherwise represent as Continuous Probability Distributions (CPDs). An NPT express the probability of each value of the child given each combination of values of its parents. However, for a variable without any parent the NPT express the prior probability for each value of the variable which usually derives from the domain expert.

In general, a BN's construction does not always follow the steps in order and often the steps are repeated several times until an optimal network has been constructed.

In [92], Verduijn et al. suggest that the construction process should start from the outcome variable and then continue by selecting the variable's Markov blanket[1] to ensure the best predictive feature subset for this outcome variable. The study applied the process recursively until a feature subset has been assessed for each variable that is to be included in the Prognostic Bayesian Network (PBN), by taking a top-down approach.

In [93] Laskey and Mahoney construct a BN by makings several 'network fragments', of which each represents different problem instances of the domain. These fragments are then combined together to model the domain with a complete BN. According to the study, a network fragment should remain separate from other fragments, and should be formed respecting the syntax and semantics of BNs.

However, the following three approaches are common to studies that focused on BNs' constructions:

- constructing a BN using the knowledge of domain experts;
- learning both the parameters and the structure of the BN from historical data;
- constructing a BN using both domain knowledge and historical data;

We discuss these approaches in more detail in Section 3.2.1, 3.2.2 and 3.2.3 below.

[1] The Markov blanket of a variable X in a Bayesian network is the set S_x of neighbouring variables that separate the variable from all other variables [92].

3.2.1 Constructing BNs based on domain experts

Many researchers have constructed BNs for applications in various domains based on the knowledge of experts [88][94][95][96][97][98][99]. In [98], the authors constructed a BN for predicting the probability of breast cancer based on a number of risk factors, using expert knowledge. They concluded that both experts and literature should be used to provide sufficient information regarding the associations between different radiological features and breast diseases.

During a BN construction process, experts also make causal assumptions based on their knowledge of the domain that is being considered. Knowing that there exists a relation between two variables, the experts must then identify the direction. A set of such directed relations forms the complete structure of the problem domain

Once the structure is complete, the experts then focus on quantifying the model with probabilities. There are a number of techniques which can assist in estimating the probabilities. Details of these techniques are given in [100][101]; we do not review them here as this thesis makes no use of probabilities elicited from experts.

The expert dependent approach to building a BN, including both structure and parameters, assumes that the experts should be able to express their knowledge and estimates precisely and accurately. However, this is known not to be the case and in particular experts from the clinical domain often struggle to articulate the knowledge needed for constructing an expert system [102]. In [103] the authors stated that since the relations in their models have been specified from the domain experts these models have the potential of being biased. Studies have also shown that the parameters estimated by experts are subject to cognitive biases [86]; a demonstration of this was made in [104].

3.2.2 Learning BNs from data

The approach of learning a BN from data can be considered to fall into two categories: the first, simpler, category uses an algorithm to learn the parameters of a BN from a

dataset, when the structure of the BN is given; the second learns both structure and parameters. We consider these in turn.

Learning parameters

The most common approach is to calculate parameters based on the relative counts of the outcomes; however, this assumes the data are complete. When the data has some missing values, a common approach parameter learning is Maximum Likelihood Estimation (MLE) [105][106]. The method uses an Expectation-Maximisation (EM) algorithm to find these maximum estimates of parameters [107]. EM is an iterative method in which iteration alters between an expectation step (E), and between a maximisation (M) steps. The E-step computes the expectation of the likelihood based on evaluation using the initial estimate of the parameters, and the M-step computes the maximum likelihood estimates of the parameters by maximising the expected likelihood that is found on the expectation step. The estimated parameters in this M-step were then used to determine the distribution of the variables in the next E-step.

Learning the structure

Various algorithms have been developed for learning the structure of a BN, and these algorithms mainly fall into two categories: search-and-score based methods and constraint based methods. The former approach considers learning as a model selection problem. Algorithms of this approach search the space of all candidate structures for the variables by using a heuristic for the one that best represents the probabilistic relationships with respect to data [108].

The latter approach considers learning as a problem that requires finding structure from the idea of independence. Algorithms of this approach take account of some predefined constraints to learn relations from data. In general, a constraint is derived by comparing the result of a test of conditional independence (CI) in the data to a threshold. The structure in which all the relations meet such constraints considers as the best for explaining dependencies and independencies with respect to the data.

Various scores have been considered for learning BNs structure using the score-and-search algorithms, such as: the likelihood score, the Bayesian information criterion

(BIC) score, the Akaike information criterion (AIC) score and the Bayesian score. The likelihood score is determined from the likelihood of data given the structure, and then the maximum of the function is used to calculate the maximum likelihood score of the BN. The Bayesian score is a measure of how well a given structure,M, fits the data, D, is defined as:

$$Score\,(M,D) = P(M|D) \qquad 3.8$$

Equation 3.8 is essentially the posterior probability of M, given the data, D. While learning, a score-and-search algorithm has to maximise the score and the Bayesian approach performs this computation using Bayes Theorem:

$$Score\,(M,D) = P(M|D) = \frac{P(D|M)P(M)}{P(D)} \qquad 3.9$$

The denominator of the equation does not help to differentiate between different structures, and therefore, in order to maximise the score the numerator needed to be maximum. If we ignore $P(M)$, then the score calculates as:

$$Score\,(M,D) = P(M|D) = P(D|M) \qquad 3.10$$

To compute $P(D|M)$, the Bayesian approach averages over all possible parameters, weighing each by their posterior probability:

$$P(D|M) = \int_{p_m} P(D|\,p_m\,,M)P(p_m|M)dp_m \qquad 3.11$$

Search-and-score algorithms are heuristic and often the basis of this is a greedy search

(GS), but other methods are also used. Daly et al. [109] give a review of the methods that have been developed for learning the structure of Bayesian networks from data.

Michalowski et al. [110] construct the structure of a Bayesian Belief Network (BBN) based on the K2 algorithm [108], which uses a greedy search method, starting from the assumption that a variable has no parents. K2 works in a loop over all the variables: at each stage it adds incrementally the parent whose addition increases the probability of the resulting network. When it becomes impossible to increase the probability of the network given the data by adding another single parent, the process stops and gives the BBN that has the maximum posterior probability on the data.

For the structure of a data-drive Bayesian network Suebnukarn et al. [111] have used the Necessary Path Condition (NPC) [112] – a constraint based learning algorithm. The motivation was to identify all the conditional independence and dependence relations between variables. The learning begins with a set of statistical tests and completes the structure through several steps. The steps are followed to establish directed links based on conditional independences as found from the tests while creating a directed acyclic graph.

3.2.3 BNs from data and domain experts

Researchers have also considered combining expert knowledge and data to construct a BN. In this method, the causal structure of a BN is derived from experts, who mostly acquire relevant knowledge from their experience or from the available literatures. Then, the parameters for this structure are learned from a dataset to complete the model construction. The method is now very popular in epidemiology. In particular, researchers prefer to construct a causal structure from the experts in advance and uses it to guide the analysis of epidemiological data afterwards [113][114]

In addition to above, there are other ways to combine the knowledge of expert with data for constructing a BN. Khan et al. [115] incorporated knowledge with data for learning parameters. They outlined parameter constraints using domain knowledge and incorporated these constraints in a separate BN model. The study data samples were

incorporated in this model to learn parameters given constraints. In [9], Gevaert et al. have specified the priors for the structure and parameters from domain experts. Each of the structure priors is essentially a probability that corresponds to each directed link from all combinations of links between two variables. The model is based on three variables, this has generated six links and the experts have therefore specified six priors for the structure. Then, these priors were assessed by the experts and changed according to the study dataset. Their approach has demonstrated that the prior probabilities about the structure of a BN, as specified from experts, could be used to guide in learning the structure considering data.

In Dekker et al.'s [7] learning approach, the domain experts have made a draft structure at first. Taking this as a starting point a search algorithm – the Markov Chain Monte Carlo (MCMC) method [116]–was later used to select the optimal structure that results in producing a higher predictive likelihood on data.

3.3 Tools for BNs

There are various software tools that are used both to construct a BN and for completing all the necessary inference calculations. Some tools are non-commercial [117][118][119][120][121][122][123][124][125][126] and others are commercial [127][128][129][130][131]. In [132] Murphy has listed all the features of each of the above software tools to help developers in deciding which one is the best in terms of the required BN and the extent of its support.

In this thesis we make use of the AgenaRisk software tool [127] for constructing all the required BNs. This is because among all the available tools, AgenaRisk has all the features that we need to conduct and complete our study. In particular, the tool:

- has a powerful graphical user interface;
- allows modellers to construct a complex and large BN model in a simple way;

- provides a Java application program interface (API) allowing both model construction and model inference to be scripted so that, for example models can be generated from a data file;

- supports hybrid BNs with continuous variables handled using dynamic discretisation.

Taking account of all the above, we judged AgenaRisk to be the best tool for our study. However, it does not yet provide supports (a) for learning the structure of a BN from data and (b) to evaluate the strength of a causal relation, in a straightforward way, but it enables us to complete both (a) and (b) by experimenting with the underlying Java code. In particular, the API of the software tool allowed me to go to any extent during the implantation phase of the learning algorithms that I proposed for my studies.

3.4 Summary

In this chapter we have described the basic concepts of Bayesian networks, including BN inference, static and dynamic discretisation methods, together with the steps needed to construct a complete BN model. We presented some examples of the approaches used to construct BN for specific applications in a number of different categories. Finally, we reviewed some BN tools and explained our selection of AgenaRisk.

Chapter 4

Bayesian Networks in the Clinical and Health Care Domain

This chapter surveys the existing applications of BNs to clinical support. Section 4.1 looks at the ways that BNs have been used in the clinical and health care domain. Section 4.2 discusses the methods that have been applied to evaluate the performance of these networks. Section 4.3 examines the use of these techniques on a sample of published applications. Finally, Section 4.4 compares the existing practices against the research objectives of this thesis.

4.1 The use of BNs for clinical reasoning

BNs are popular tools for modelling risk and uncertainty. They have been successfully applied in diverse fields within the clinical domains. We review a number of such studies in this section by categorising them according to the tasks that they support.

4.1.1 Diagnosis

Usually a clinician performs diagnosis of the diseases of a patient from the symptoms, so that 'diagnostic reasoning' has become the general term for inferring information about causes (e.g. a disease) from observed consequences (e.g. symptoms of illness). However, diagnosis is never certain; even after taking all the symptoms into account the clinician remains unsure – to some extent – about the true condition of the patients. To avoid mis-diagnosing the patient, this uncertainty must be considered. BNs have been

found to be a suitable framework for representing uncertainty in clinical problems. This section reviews a number of studies that use BN models for diagnosis.

Alvarez et al. [133] used a BN for diagnosing pyloric stenosis, and concluded that this model is better for detecting the disease than clinicians and reduced the number of recommendations for ultrasounds. Kline et al. [134] investigated a BN to estimate patients with a low pre-test probability (between 0% to 2% ranges), for helping clinicians making decisions about diagnostic tests for embolisms. Also for the diagnosis of pulmonary embolism, Luciani et al. in [135] described the use of BayPAD – a probabilistic expert system –, for improving the accuracy of diagnosis. Haddawy et al. described a BN in [136] to diagnose and select the procedure for patients with suspected gallbladder disease.

Watt and Bui [103] constructed a BN from a combination of domain expertise, data and the findings of a literature review. The BN was for the prediction of knee osteoarthritis (OA); the authors concluded that the model appeared to be effective in identifying the symptoms that correspond to the presence of OA in the knee of a patient, but the knowledge taken from the experts can introduce bias into the relations modeled in the BN.

A number of BNs have also developed to estimate an individual risk of diseases [98][137][138][139][140][4]. In addition to the detection of a disease, BNs have also been developed and used to identify patients to decide guideline eligibility for disorders such as asthma [141], and pneumonia [142].

4.1.2 Classification

The aim of a classification model (also known as a classifier) is to categorise cases into one of the n mutually exclusive outcomes. Most of the BNs that have been developed to perform diagnosis tasks can be regarded as Bayesian classification models, but classification is more general since most diagnostics models classify patients into one of two categories.

Liu et al. [143] applied a Bayesian classification model to differentiate between benign and malignant Thyroid Nodules by using sonographic features. The model includes features that are known to be the predictor of malignancy as well as factors that significantly influence in observing malignant nodule, such as age and gender. The variable *Age* had two discrete states: < 50 years and >=50 years. Because of this it was not possible for the classifier to make any distinction between a patient aged 25 and a patient who aged 49, or between a patient aged 50 and a patient aged 85.

In [144] Chattopadhyay et al. present a naive Bayesian classifier to predict presence/absence of dental diseases from a set of parameters indicating toothache among study patients. Lee et al.[145] proposed a Bayesian classifier to identify malignant renal cyst. They considered the findings from experts as a risk to the performance of their BN classifier for predicting malignant renal cysts.

Further, Wu et al. [146] introduced a Bayesian classifier in an attempt to develop a computer aided diagnosis of Cerebral Aneurysm, Burnside et al. [147] proposed a model for predicting breast cancer, and Blanco et al. [148] applied a Bayesian classifier for predicting the survival rate within the first 6 months after Transjugular Intrahepatic Portosystemic Shunt (TIPS) placement. While the nature of these models is binary, there are also other Bayesian classification models which consider multiple, i.e. more than two, outcomes during a classification task.

Stojadinovic et al. [149] developed a Bayesian classifier for predicting breast cancer risk in younger women (age < 40 years). The classifier is used to categorise each woman's breast cancer into one of three levels: benign, malignant and pre-malignant. The study was based on 591 women and the dataset contained information about age, ethnicity, clinical history, hormonal information, breast density and size, risk estimate (from a Gali model), results of imaging and biopsy. However, the study demonstrated that the recruitment process of study subjects isn't really a straightforward task and had to be managed carefully.

4.1.3 Prognosis

BNs that predict the outcome of an event in the future are known as Prognostic Bayesian Networks (PBNs). With a focus towards the future, uncertainty is even more important in PBNs than in the BNs developed to support diagnoses. The structural representation of a PBN can model the time-variant nature of a problem to allow predictions to be performed in stages. At each stage, the model uses all the information that has become available at the time of the stage to perform the prediction.

Verduijn et al. [150] developed a PBN for the domain of cardiac surgery using variables identified from three stages: pre-assessment, operation and recovery (postoperative). Taking account of the complications that occur during the operation and postoperative stay of a patient, the PBN predicts mortality. The study is based on data of 10147 patients who had cardiac surgery in their selected hospital between January 1998 and November 2004. The PBN included multiple outcome variables, namely *hospmort*, *ORmort* and *postORmort*, which respectively represent the hospital mortality, operative mortality and postoperative mortality of a patient. By dividing the dataset into two subsets (a) training and (b) testing, the authors have learned the model based on (a) whereas used (b) to validate its performance. Further, the predictive ability of the PBN was compared with another network (Figure 4-1) that maintains the time-dependent ordering of the variables of the PBN learnt using the Minimum Description Length (MDL) principle.

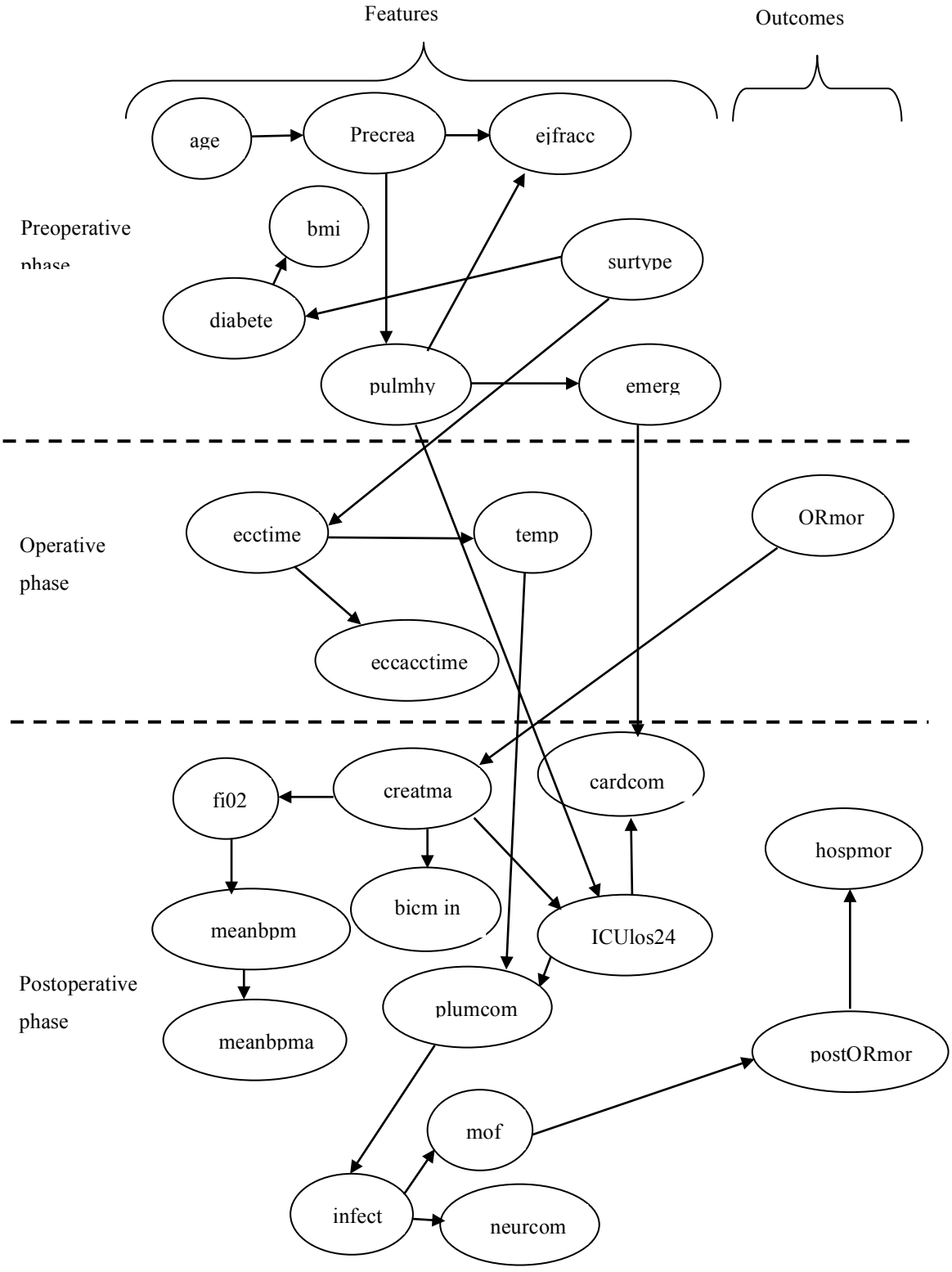


Figure 4-1 MDL based PBN to demonstrate the time dependent ordering of variables

Other BNs developed in the clinical domain to predict outcomes include BNs for assessing morbidity and mortality in patients who had coronary surgery [10], liver transplantation [96], injury [151], suffer with sickle cell disease [152] and lung cancer [153], for predicting the outcome of endodontic treatment [111] and the response to bronchodilator medication [140].

In [140] a hybrid Bayesian network predicts Bronchodilator response (BDR) by taking account of simultaneous associations and interactions between variables. The model was based on data from a cohort of 308 Caucasian Childhood Asthma Management Program (CAMP) – a clinical trial – subjects, and constructed with a greedy search method. The predictive ability of the model was determined through a fivefold cross-validation method (see Section 4.2.1). According to the authors of the study, some variables that may contribute significantly to BDR may have remained unidentified and were therefore absence from the search process, and this was the reason for obtaining a disappointed performance from the model.

BNs have also been used to obtain predictive values to provide assurance of better allocation of resources and management of health care. For instance, in [154] the authors developed BNs to monitor patients and aid in drug therapy. Their system incorporated both general population data and incoming patient data to provide patient-specific models. The BBN-RPP model in [110] provides probabilities over the values (met or delayed) of the outcome variable LOS (length of stay) and helps to assess the impact of a patient's outcomes and activities on his length of stay at a hospital after surgery. The triage model in [97] estimates the probabilities of different decisions in order to determine if a patient coming to an emergency department (ED) actually requires hospitalisation. Further, [155] develops a model to predict the length of stay of patients as a tool to help the management of hospitals to improve better geriatric health care.

This section has shown that BNs have been developed for a variety of application with several different styles of reasoning. But in all cases, the aim of the BN is primarily to update the probability distribution of one, or a small number, of outcome variables. As we will see in later chapters, the focus changes when we attempt to use a BN to analyse data gathered by observing process in operation.

4.2 Performance of Bayesian nets

Researchers have mostly assessed the predictive performance of a BN by examining three abilities: 1) discrimination ability 2) generalisation and 3) calibration. In this section, we review the methods to assess each of these three abilities. The following section looks at the way these methods have applied to the clinical BN projects reviewed in Section 4.1.

4.2.1 Discrimination or classification ability

The discrimination ability of a BN model measured how well the model can distinguish between two or more classes. Either a model can be tested using a separate dataset for testing (as opposed to training the model) or a cross validation can be used. In general, an N-fold cross validation divides the study dataset into N equal subsets and from these N-1 datasets are used to learn BNs accordingly and the Nth dataset is used to test the BN. This is repeated for N separate tests, so that all data points have been used both to train and to test (though never together). The performance of the N tests is averaged.

The results of a test separate cases into four categories, namely: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The names assume the outcome variable is Boolean, but the process can be used for any 2–state. Using these categories, the measures shown in Table 4-1 are derived for evaluating the network.

These measures assume an ‘operating point’ for the classifier. The operating point is determined by the probability threshold used for classifying the outcome into the two categories: the obvious value of 50% is not necessarily the best; instead different operating points trade-off between sensitivity and specificity.

Table 4-1 Measures for evaluating the ability of a Bayesian network

Performance measure	**Definition**
Sensitivity (Se)	$Se = \frac{TP}{TP+FN}$
Specificity (Sp)	$Sp = \frac{TN}{TN+FP}$
Positive Predictive Value (PPV)	$PPV = \frac{TP}{TP+FP}$
Negative Predictive Value (NPV)	$NPV = \frac{TN}{TN+FN}$
Correlation Coefficient (CC)	$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+sFP) \times (TP+FP) \times (TN+FN)}}$

To show the discriminating ability of a BN at different operating points, a Receiver Operating Characteristic (ROC) curve can be used. To obtain an ROC curve, one evaluates the performance of the model with a number of possible threshold values. The ROC curve then plots the sensitivity against the specificity for different threshold values to show the overall performance of the model. The area under the ROC (AUROC) is used as a measure of the model's performance. It ranges from 0.5 to 1, where the lower value indicates that the model performs no better than random guessing whereas the upper value indicates that the model performs its task perfectly.

Rather than creating a test dataset from that used to develop a BN model, many have used an independent dataset for testing. Figure 4-2 is the ROC curve generated in [1] when the authors test their BN – that has been developed from a cohort of 9349 patients, with an independent dataset containing 992 patients.

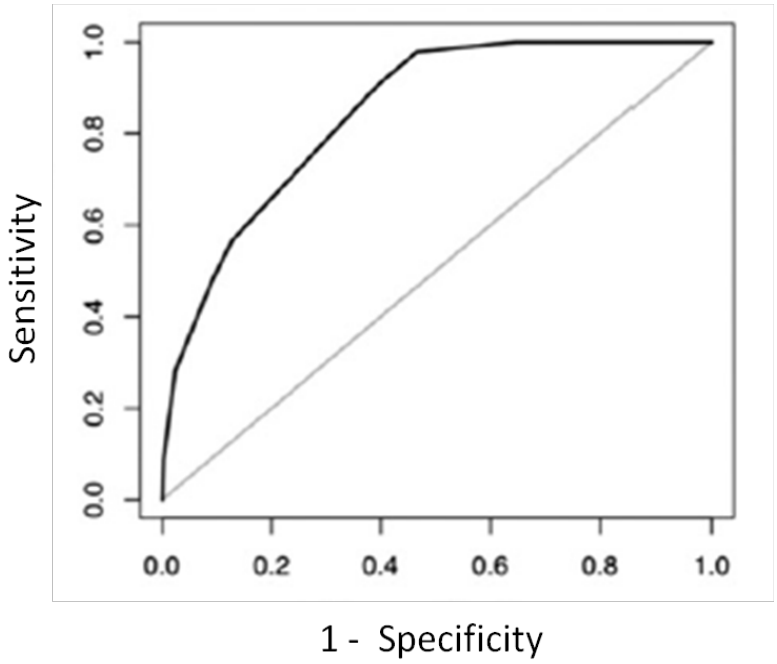


Figure 4-2 The ROC curve from the independent dataset in [1]

While an AUROC reveals how accurate a model is for discriminating cases into two outcomes, the measures mostly used for a BN that concerns multiple (more than two) outcomes are its sensitivity, specificity and positive predictive values. For instance, the BN in [97] models the process followed by a triage nurse to handle patients arriving in an Emergency Department (ED), with one of four possible triage decisions: (1) call 911, (2) go to emergency room, (3) visit doctor in the next two business days and (4) self care. To assess the performance of the model, the study only measures the sensitivity and specificity of the model for each decision.

Miettinen et al. [156] constructed a number of BNs from several scoring functions and have applied them to classify a case into one of the six otoneurological diseases. Each of the models was tested using data in a ten-fold cross validation and to assess this method for representing the dependence relations between the identified attributes of otoneurological data, they calculated the average of the sensitivities, positive predictive values and accuracies, which is then compared with the result of the tests performed with neural networks.

Instead of using any of the techniques described above Reynolds et al. [157] have calculated error rates to assess the predictive abilities of the BN classifier they developed. In particular, they have applied the BN on 46 tumours to classify each into one of seven classes, choosing the one with for which the BN predicts gives the highest probabilities; they then calculate three different errors to measure the performance of the classifier.

4.2.2 Generalisation ability

The generalisation ability of a BN model is generally measured by testing the model with a test dataset, separate from the training dataset. It is essential to consider generalisation to know how well the model will perform when it is used for cases that haven't been used in its development. In addition to the use of a test dataset, there are also other ways in which the generalisation ability of a BN prediction model can be assessed. In some studies, after obtaining the results from a test, the authors have further compared them with the values processed by (a) another model and (b) the domain experts.

In most of the cases seen, the 'another model' used is essentially a model that is currently in use for providing probabilistic estimations and the BN is learned completely from the study data. If there is no model currently in use, the relevant probabilities (for e.g. prognosis) can be estimated by the domain experts.

4.2.3 Calibration ability

The calibration ability describes how close the estimates of the model are to the true underlying probabilities [158]. A common measure is the Hosmer-Lemeshow (HL) goodness-to-fit statistic [159], which summaries a calibration curve. In order to calculate the HL goodness-to-fit statistic, the cases are placed into groups by first predicting the outcome for each case from the model, and then ranking the cases according to the predicted probabilities. The statistic is then derives from both the predicted probabilities and actual results of each group. To explain how well a model fit

to the data, many studies use the P-value of the derived HL goodness-to-fit statistic. Despite this widespread practice, Biagioli et al. [10] have stated that for a large dataset the use of P-value can lead to misleading conclusions.

4.3 Use of the performance measures in practice

In Section 4.1 we categorised a number of studies according to the task that the BN they develop supports. In this section, we tabulate the measures used for evaluating the performance of 32 published BNs, covering diagnosis (Sections 4.1.1), classification (Section 4.1.2) and prognosis (Section 4.1.3). The results are in Table 4-2.

Table 4-2 Evaluation methods and measures for BNs discussed in sections 4.1.1 to 4.1.3

Authors	Evaluation method					Performance measure						
	Cross validation	Expert's validation	Other models	Test dataset (no cross validation)	Independent dataset	AUROC	Se	Sp	PPV	NPV	HL statistic	Other
Alvarez et al. [133]	No	Yes	No	No	No	Yes	Yes	Yes	No	No	Yes	No
Kline et al. [134]	No	No	No	No	No	Yes	No	No	No	No	No	A test threshold computed from Pauker and Kassirer [160]'s formula
Luciani et al. [135]	No	No	Yes	No	No	No	Yes	Yes	No	No	No	No
Haddawy et al.[136]	No	Yes	No	No	No	No	No	No	No	No	No	No
Burnside et al. [98]	No	Yes	No	No	No	Yes	Yes	Yes	Yes	No	No	No

Cruz-Ramírez et al. [137]	No	No	No	Yes	No	No	Yes	Yes	Yes	Yes	No	No
Maskery et al.[161]	No	Yes	No	No	No	No	No	No	No	No	No	Brest pathology literature for evaluation
Watt et al.[103]	Yes	No	Yes	No	No	Yes	Yes	Yes	No	No	No	No
Sanders and Aronsky [139]	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Positive likelihood ratio and negative likelihood ratio
Himes et al. [140]	Yes	No	No	No	Yes	Yes	Yes	Yes	No	No	No	Standard error for the AUROC
Charitos et al.[4]	No	No	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No
Sanders and Aronsky [141]	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Kline et al.[134]	No	Yes	No	Yes	No	No	No	No	No	No	No	95% Confidenc-e Interval
Aronsky et al. [142]	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Positive likelihood ratios, negative likelihood ratios and test effectivene-ss
Stojadinovic et al. [149]	Yes	No	No	No	No	Yes	No	No	No	No	No	No
Lee et al. [145]	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	No	No
Liu et al. [143]	No	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	No
Chattopad-hyay et al.[144]	No	No	Yes	Yes	No	Yes	No	No	No	No	No	No
Wu et al[146]	Yes	No	Yes	No	No	No	No	No	No	No	No	Analysis of the error rate and the image recognition rate
Burnside et al.[147]	No	No	Yes	No	No	Yes	No	No	No	No	No	No
Blanco et al.[148]	Yes	No	Yes	No	No	Yes	No	No	No	No	No	Calculated the Brier score to

												estimate the classifier calibration
Verduijn et al. [150]	No	No	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No
Biagioli et al. [10]	No	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No
Burd et al. [151]	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No
Sebastiani et al.[152]	No	Yes	No	No	Yes	Yes	Yes	No	No	No	No	Error rate
Jayasurya et al.[153]	No	No	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No
Suebnukarn et al. [111]	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	No	No
Himes et al[140]	Yes	No	No	No	No	Yes	Yes	Yes	No	No	No	No
Berzuini et al. [154]	No	Yes	No	No	No	No	No	No	No	No	No	No
Michalows-ki et al.[110]	No	No	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No
Sadeghi et al. [97]	Yes	Yes	No	No	No	No	Yes	Yes	No	No	No	Chi-square test
Marshell et al.[155]	No	No	Yes	Yes	No	No	No	No	No	No	No	Chi-square test

4.3.1 Critiques of the performance methods

Table 4-2 presents various existing methods to evaluate the performance of a BN model. The table shows that in many cases the accuracy of a model constructed from a training dataset has been assessed on a test dataset. While some studies used an independent dataset to test a BN model, others have mainly selected the test dataset from the data used during the development, perhaps using a cross validation.

Ohno-Machado and his colleague [162] show concern about separating a dataset into a test dataset and a training dataset. Indeed, there are various studies within the clinical literature in which the authors conclude that a significant difference between the sizes or qualities of a training dataset compared with the test dataset is the foremost reason for the BN constructed in the study not being applied in practice [152][163][9][164]. In [161] Maskery et al. discussed that using train-test datasets to validate a BN's ability is only suitable when there is only one outcome for the problem of interest. It follows that

the methods of cross-validation are not appropriate for validation if a BN provides an adequate representation of observed data.

Further, the table shows that the studies have mostly calculated the area under an ROC curve to assess a BN's accuracy in making predictions. The ROC curve for a BN results from the sensitivity and specificity of the model. In [165] Yet et al. argue that the common measures used for the accuracy of a BN (which they refer as a scoring system in the study) are inappropriate for making the model acceptable in the clinical domain.

4.4 Summary

This chapter discussed the existing applications of BNs to medical support. First it investigated the ways that BNs have been used in the clinical and health care domain. Then it discussed the methods that have been applied to evaluate the performance of these networks. Following that, it examined the use of these techniques on a sample of published applications.

Chapter 5

Case study: Evaluating Treatment Selections for Patients with Cancer – Initial BN's Construction

In [166] Victor Ogunsanya, a researcher from the Risk and Information Management (RIM) research group at Queen Mary, University of London, used a case study to investigate a parameter learning approach using data. His study required a small BN to be constructed, representing relations between three variables from the domain. A model of this size is insufficient of the domain. This chapter uses the same case study but extends the analysis- a new dataset is collected and analysed to identify more variables of the domain. Relations between these variables are determined using the knowledge of the expert Professor Hemant Kocher. These new variables and their relations help to enlarge the size of the BN proposed in [166].

The expert who the research group consulted to retrieve information regarding the structure of the expert BN is Professor Hemant Kocher. He joined in Barts Cancer Institute in 2005 as a Senior Clinical Lecturer. Some of his academic achievements are a National Clinician Scientist Fellowship by the National Institute of Health Research (UK), an MD from Kings College London in 2003 and specialist clinical training in hepato-biliary-pancreatic surgery in 2005. His research interests include: liver cancer, pancreatic cancer, surgery of the gall bladder and bile duct, the liver and the pancreas, developing new treatments, risk assessment and quality improvements. The case study discussed in this thesis was a collaborative project and uses the data collected from Professor Hemant Kocher's research group at the Barts and the London HPB (HepatoPancreaticoBiliary) Centre at the Royal London Hospital of Barts NHS Trust.

A number of consultation sessions were held to and information regarding the followings was extracted from judgements made by the expert:

- How many of the identified study variables adequately represent the MDT process?
- Which variables relate with each other?
- What information are there to describe a possible relationship between two variables?

The HPB Centre at the Barts and the London NHS Trust (BLT) is a specialist centre offering treatment and care to patients with biliary, pancreatic and liver diseases with the help of teams of specialists consisting of oncologists, surgeons, physicians, pathologists, radiologists and palliative care specialists, who form a Multi-Disciplinary Team (MDT). The data come from the operation of the MDT meetings, at which treatment decisions are made; these processes are introduced in Section 5.1. The following two sections give an overview of the data and how the data were used to select variables in a BN model of the MDT outcome. This model is then used in both Chapters 6 and 7 as a case study for novel techniques for the evaluation and analysis of observed data, which are the main contribution of thesis.

5.1 Multidisciplinary team meetings

A multidisciplinary team meeting (MDT) is an important decision-making forum in medicine, especially in oncology. The objective of the HPB's MDT is to generate comprehensive and coordinated care plans for patients with cancer or suspected cancer. Previous studies of the use of MDT meetings in oncology have shown that patients managed by such meetings have better survival outcomes [167], shorter waiting times [168], and benefits from a more robust treatment selection process [169], compared to those managed without formal multidisciplinary discussions.

Case Study: Evaluating Treatment Selections for Patients with Cancer – Initial BN's Construction

Suspecting cancer as the possible disease, GPs in the BLT's district initially refer their patients to one of the twelve general hospitals in the district. These patients are then referred to the BLT HPB centre where their cases are scheduled for further discussion in the MDT meetings held at the HPB centre. The meetings review imaging and the results of other tests, reaches a diagnosis and recommends one or more treatment options.

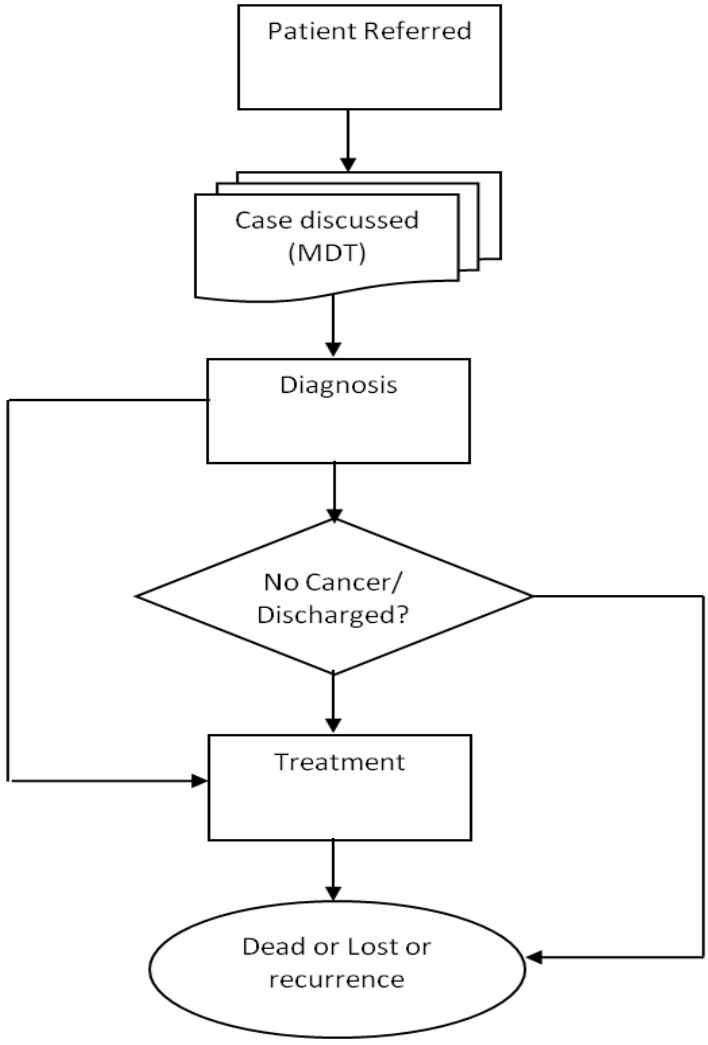


Figure 5-1 The process of recommending treatments to patients referred to the MDT

If the available information for a particular patient is insufficient for the MDT to make a treatment recommendation at the first meeting, the case is discussed in another meeting. A patient's case can also be discussed in multiple MDT meetings if it is thought to be really complex. As well as choosing a treatment route, other outcomes are also possible.

For example, when the MDT reaches the conclusion that the probably diagnosis for a patient is a non-cancerous disease, the patient is discharged to other specialists. Figure 5-1 represents the process of recommending treatments to patients referred to the MDT, using a flow chart.

5.2 Data

The source data used in this study contains records of patients who suffered with cancer or suspected to have cancer and were discussed at the MDT meetings held at the HPB centre between 2005 and 2009. During this period, meetings were usually held each week and the staff initially wrote up details of the patients discussed in a word document. Data were then extracted from these word documents and stored in an excel data file. In total, the file contains records of 2074 patients. Because of the way the data were captured, the relationship between data fields and variables in the BN is not straight forward. We first describe the raw data and then the way this is used to create a BN.

The data file included details of seven attributes: data of birth, data of discussion (this is when a case is first discussed after being considered for an MDT meeting), the number of meetings held, affected organ, type (i.e. the severity stage of a diagnosed cancer), diagnosis, and treatment option. Although the data file mostly includes only one value for each attribute, in some cases it contains multiple values of treatments to suggest that the patients were recommended multiple treatments at the MDT meetings.

A diagnosis of a patient's cancer is made using 44 possible values as presented in Table 5-1. Each of these values corresponds to one of the possible categories of cancer that a patient may have depending on the organ that had cancer. There are fifteen possible categories if the cancerous organ is the pancreas and five possible categories of cancer if the organ is the gallbladder. The numbers of possible categories of cancer when the organs are the liver and the bile duct of the patient are 14 and 10, respectively.

Table 5-1 Possible categories of cancer for each organ

Organ	Diagnosis
Pancreas	PDCA (Pancreatic ductal adenocarcinoma)
	PNET (Pancreatic neuroendocrine tumour)
	IPMN (Intraductal papillary neoplasia)
	MCN (Mucinous Cystic Neoplasms)
	SCN
	CP
	AP
	Cyst
	Stone
	Other
	Pseud (Pseudopapillary tumour)
	Mets (Metastases)
	Pancreas divisum
	SPN (Solid pseudopapillary neoplasm)
	Unknown
Liver	HCC (Hepatocellular carcinoma)
	CRC Mets (Colorectal cancer liver Metastasis)
	OO Metastasis
	Fatty Liver
	Cirrhosis
	Adenoma
	FNH (Focal nodular hyperplasia)
	Cyst
	Hydatid Cyst
	Haemangioma
	Liver abscess
	Hepatobiliary cystadenoma
	Angiomyolipoma
	Other
Gallbladder	Cholecystitis
	Stone
	Malignancy
	Adenoma
	Other/Unknown
Bile duct	Cholangiocarcinoma
	Bile duct stricture
	Iatrogenic stricture
	Iatrogenic Leak
	PSC (primary sclerosing cholangitis)
	CBD Stone (Common bile duct stone)
	MS (Mirizzi's Syndrome)
	Inflammatory strictures
	Cholangitis
	Other/Unknown

The MDT meetings recommend six treatments to the patients discussed. These treatments are: chemotherapy, radiotherapy, surgery, palliative care, intervention radiology and watchful waiting. In some situations, rather than recommending a single treatment the meetings may recommend more than one treatments. However, it is not always the case that a treatment has been recommended to a patient. In these cases, the data file contains an entry of either (a) lost, (b) death, (c) discharged or (d) no cancer to indicate one of the following:

- The MDT was unable to trace the patient to recommend a treatment.
- The patient died before receiving a treatment from the MDT.
- The patient has been discharged.
- The type of the diagnosed cancer was benign.

In addition to the main data file, additional summary statistics about the MDT were made available for this study. Table 5-2 presents the summary statistics covering (a) total number of cases discussed in each year (b) average time taken per MDT meeting and (c) average time taken per case for the year considered. This reveals that the average time taken for each patient steadily decreased over the study period.

Table 5-2 Summary statistics of the MDT meetings for the study period

Year	Discussed patients (total)	Average time per MDT meeting (in minutes)	Average time per patient (in minutes)
2005	7	90	12.52
2006	10	110	11.00
2007	15	120	8.00
2008	20	120	6.00
2009	30	105	3.30

5.3 Selection of variables

This initial excel data file appeared to have inconsistent records for some patients which has caused a significant problem during the selection of variables. So some refinements are made to the data file based on evidence of inconsistencies. In particular,

- The records of the 81 patients who had no information of date of births in the data file have been deleted. Then, the records of 6 more patients were deleted because of an obviously wrong date of birth value. At this stage, the data file contains records of 1987 patients in total. We refer to this dataset as the 'Meeting data'.

- The records of 89 patients were deleted for having no information about the cancerous organs. Similarly, the records of a further 347 patients were deleted. This is because, for 187 of these patients the record of cancerous organs was unknown and for 160 patients, the cancerous organ is not any of the four organs (the pancreas, liver, gallbladder and bile duct) considered at the MDT meetings. After deleting these patients, the data file has records of 1551 patients in total; we refer to this dataset as the 'Organ data'.

- A further 25 patients had no information on the disease type and these records were also deleted. This brings the total number of patients to 1526. The data file of the records of 1526 patients is referred to as the 'Organ-Type data'.

- A further 50 patients had no or a wrong diagnosis value. The records of these patients were deleted and the data file now of 1476 patients is referred to as the 'Diagnosis data'.

- Finally, 30 more patients with a wrong treatment option have been identified and deleted from the data file. After this last refinement, a data file with 1446 patients is obtained. This file is referred to as the 'Treatment data'.

These deletions from the initial data file have enabled us to find the maximum number of study patients with complete records of each model variable. We selected seven discrete variables: *Age*, *Year*, *Number of meetings*, *Organ*, *Type*, *Diagnosis* and *Treatment.* These variables are described in detail below, noting how the original range of values has been mapped into the BN model.

Age

The variable represents the approximate age of a cancer patient at the time of the first MDT discussion: this is derived from his date of birth and his case discussion date. This age is discretised to one of the seven groups: *Under 46*, *46 – 54*, *>54 – 60*, *>60 – 66, >66 – 71*, *>71 – 77*, and *77+*. Each age group holds an equal number of patients and this was around 13% of the study patients.

Year

The variable represents the year when a case is first discussed after being considered for a MDT meeting and is selected from the date of discussion. The variable contains five values: *2005*, *2006*, *2007*, *2008* and *2009*.

Number of meetings

The variable represents the number of MDT meetings held per case within the period. For most of the cases, the recommendations of treatment resulted from just a single meeting. However, for some cases a recommendation required multiple meetings and the maximum of them was 11. The values defined for this variable are: *1*, *2*, *3*, *4 or more.*

Organ

The variable represents the cancerous organs and has four corresponding values: *Pancreas*, *Liver*, *Gallbladder* and *Bile duct*.

Type

The variable represents different cancer stages. It has three values: *Benign*, *Malignant* and *Unknown*. Note that here the value *Malignant* represents both malignant and the borderline cancers.

Diagnosis

The variable represents various cancer diagnoses. Based on the cancerous organs and their possible types, the full set of 44 possible diagnosis categories were grouped into the following eight non-overlapping values:

- *BP* = a benign cancer in the pancreas,
- *BL* = a benign cancer in the liver,
- *BGB* = a benign cancer that develops either in the gallbladder or in the bile duct,
- *MP* = a malignant cancer in the pancreas,
- *ML* = a malignant cancer in the liver,
- *MGB* = a malignant cancer that develops either in the gallbladder or in the bile duct,
- *Unknown* = an unknown type of cancer,
- *Multiple* = multiple types of cancers.

Treatment

This variable represents the treatments that MDT meetings recommend to the study patients. For some patients during a meeting experts have recommended more than one treatment options and thus, to ensure the mutual exclusive property for the variable a value, *Combination*, is added. A treatment that was always recommended jointly with other treatments was radiotherapy. Those that were recommended as a single treatment option correspond to the following five values: *Chemotherapy*, *Palliative*, *Surgery*, *Intervention radiology* and *Watchful waiting*. Finally, we define another value, namely *None*, to represent cases in which no treatment recommendation was made, and this ensures the mutually exhaustive property for the variable.

Tables 5-3 to 5-6 present the distribution of the values of all study variables. For each variable the number of patients is calculated based on the stage of its selection.

Table 5-3 The distribution of values of the Age, Year and Number of meetings for patients in the 'Meeting data'

Variable	States	Percent of patients (n)
Age	Under 46	13% in each group (260)
	46 – 54	
	>54 – 60	
	>60 – 66	
	>66 – 71	
	>71 – 77	
	77+	
Year	2005	23% (458)
	2006	11% (222)
	2007	26% (520)
	2008	32% (631)
	2009	7% (156)
Number of meetings	1 = Single	72.5% (1441)
	2 = Twice	25% (492)
	3 = Thrice	2% (45)
	4 = Four or more times	0.5 (9)

Table 5-4 The distribution of the values of the Organ and Type for patients in the 'Organ-Type data'

Variable	Values	Percent of patients (n)
Organ	Pancreas	38% (576)
	Liver	43% (658)
	Bile duct	14% (210)
	Gallbladder	5% (82)
Type	Benign	37% (567)
	Malignant	60% (916)
	Unknown	3% (43)

Table 5-5 The distribution of values of the Diagnosis for patients in the 'Diagnosis data'

Variable	States	Percent of patients (n)
Diagnosis	BP = a benign cancer in the pancreas.	10% (153)
	BL= a benign cancer in the liver.	8% (113)
	BGB = a benign cancer that develops either in the gallbladder or in the bile duct.	6% (90)
	MP = a malignant cancer in the pancreas.	21% (314)
	ML = a malignant cancer in the liver.	29% (424)
	MGB = a malignant cancer that develops either in the gallbladder or in the bile duct.	10% (154)
	Unknown = an unknown type of cancer.	13% (191)
	Multiple = multiple types of cancers	3% (37)

Table 5-6 The distribution of values of the Treatment for patients in the 'Treatment data'

Variable	Values	Percent of patients (n)
Treatment	Chemotherapy	6% (91)
	Combination	13% (192)
	None	27% (384)
	Palliative	23% (262)
	Surgery	18% (262)
	Intervention radiology.	3% (45)
	Watchful waiting	10% (144)

Not all the variables in the BN model come directly from the MDT dataset. Further investigation identified more variables from available additional information. The investigation produced two variables *Time per patient*, which represents the average discussion time per case, and *Total discussion time,* which represents the total time spent per MDT meeting. The variable *Time per patient* is determined from the summery statistics stated in Table 5-2 and the variable *Total discussion time* is estimated from the values from *Time per patient* and *Number of meetings*.

5.4 The structure

The model structure is determined by identifying four relation types between the variables, namely 1) illegal, 2) definitional, 3) impossible and 4) hypothetical. We discussed with a clinician who has sufficient experience about the way treatments were recommended through MDT meetings to identify links corresponding to each of the four relation types. The links are presented in Table 5-7.

Table 5-7 Relation types that exist between the model variables

From variable \ To variable	Age	Year	Number of meetings	Organ	Type	Diagnosis	Treatment	Time per patient	Total discussion time
Age	-	*	*	√	√	*	√	*	*
Year	√	-	√	√	√	*	√	→	*
Number of meetings	*	*	-	*	*	*	√	*	→
Organ	*	*	*	-	*	→	*	*	*
Type	*	*	*	*	-	→	*	*	*
Diagnosis	*	*	√	*	*	-	√	*	*
Treatment	*	*	*	*	*	*	-	*	*
Time per patient	*	*	*	*	*	*	*	-	→
Total discussion time	*	*	*	*	*	*	*	*	-

Symbols used for representing the relation types:
*** - Illegal**
→ - Definitional
- - Impossible
√ - Hypothetical

Illegal relations

The 'illegal relation' classification was chosen when introducing the link into the model would not have any reasonable interpretation. According to the expert, a link from *Age* to *Year* is an illegal relation as the year in which the patient was discussed depends only on the year when the GP referred the patient to the hospital. Similarly, a relation between *Age* and any of the four variables: *Number of meetings*, *Diagnosis*, *Time per patient*, and *Total discussion time* is also thought to have no reasonable interpretation, and hence, such a relation is decided as an illegal relation for this study.

Definitional relations

Although our interest is to a model based experts' judged causal relations, the variables that are determined by the values of other model variables produce some definitional relations. There are three sets of definitional relations: a) those that exist between the variables *Diagnosis*, *Organ* and *Type*, b) those that exist between *Year* and *Time per patient* and c) the relation that exists between *Number of meetings*, *Time per patient,* and *Total discussion time.* In particular,

- A diagnosed cancerous organ and its type, i.e. severity stage, jointly indicate the diagnosis for the patients. For example, if the cancerous organ is pancreas and its cancer type is benign then the diagnosis that is determined for the patient is BP (benign pancreas). All known diagnoses are determined using this approach, but when the type of cancer remains unknown we determine the probability of an unknown diagnosis. Due to these features the relations mentioned in (a) are taken as definitional. The Bayesian network fragment constructed from these relations is shown in Figure 5-2.

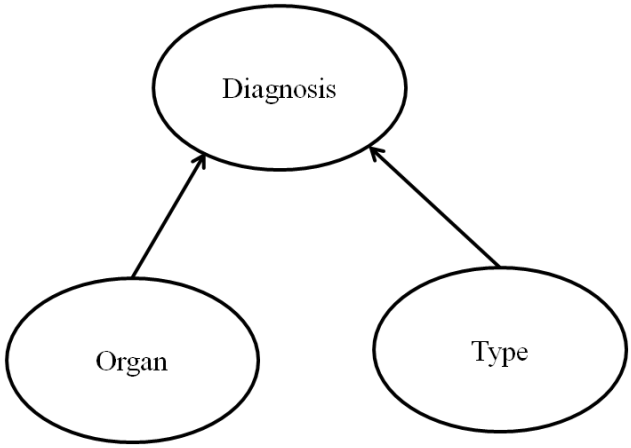


Figure 5-2 Bayesian network model fragment constructed with links based on (a)

- The average time per case per meeting in a year provides indication of *Time per patient*. The Bayesian network fragment constructed from this relation is shown in Figure 5- 3.

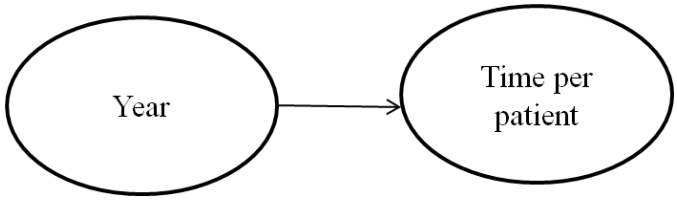


Figure 5-3 Bayesian network model fragment constructed with a link based on (b)

- The number of meetings held for a case discussion and the time that the MDT takes per case per meeting jointly indicates the total time spent in each MDT meeting. For the variables: *Number of meetings* and *Time per patient,* we therefore, define the value of *Total discussion time* using the formula stated in Equation 5.1 and take the relation mentioned in b) as definitional.

$$Total\ discussion\ time = Number\ of\ meetings * Time\ per\ patient \qquad 5.1$$

- The Bayesian network fragment constructed from the links is shown in Figure 5-4.

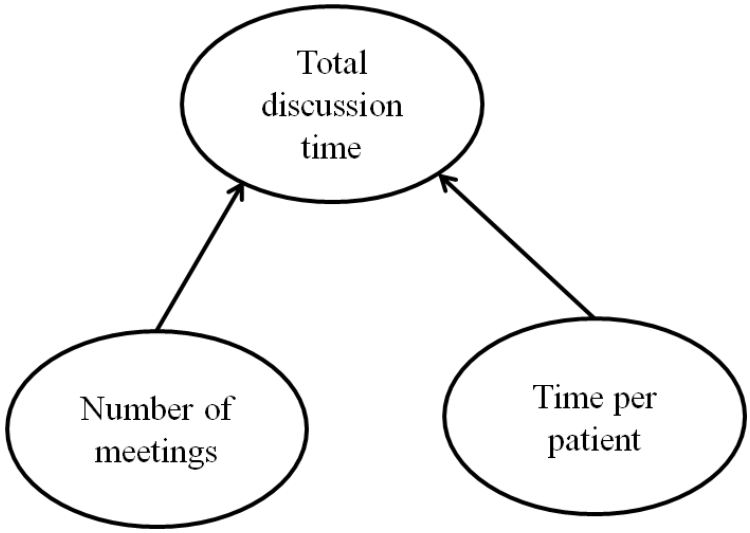


Figure 5-4 Bayesian network model fragment constructed with a link based on (c)

Impossible relations

A variable cannot have a relation to itself; such a relation has therefore been classified as impossible.

Hypothetical relations

Each hypothetical relation is one that the domain expert thinks corresponds to a causal explanation. Since the definitional relations are now known, we can explain the hypothetical relations taking account of the following four variables:

- *Age*: the age of a patient helps to determine about the cancerous organ, type of the disease and recommended treatment. The hypotheses presented in Table 5-8 reflect the above assumption of the expert. Here the variables relating to each hypothesis are presented as variable 1 and variable 2.

Table 5-8 Hypotheses for the relations between other variables and Age

Hypothesis	Variable 1	Variable 2	Description
1	Organ	Age	Changes have occurred in cancerous organs according to the age of cancer patients.
2	Type	Age	Changes have occurred in cancer types according to the age of cancer patients.
3	Treatment	Age	Changes have occurred in recommended treatments according to the age of cancer patients.

- *Year*: the year of the case discussion may influence the cancerous organ, type of the disease, treatments recommended to the patient, and the number of meetings held for the case.

  Although a hypothetical relation should corresponds to a causal relations that exists between two variables, according to our expert a relation that concerns the case discussion year is confusing. In this study, the variable *Year* represents a confounding factor, for example, a factor that represents improvements in the technologies used for detecting cancers over the period. The confusion arises because this factor is not known and therefore has not been recorded in the data. Similar possible relations may also influence the cancerous organ, type of the disease, recommended treatment and the number of meetings held to discuss the case. The hypotheses for explaining such hypothetical relations between *Year* and other variables are presented in Table 5-9.

Table 5-9 Hypotheses for the relations between other variables and Year

Hypothesis	Variable 1	Variable 2	Description
1	Organ	Year	Changes have occurred in cancerous organs over the years.
2	Type	Year	Cancers with a benign diagnosis have increased over the years.
3	Treatment	Year	Changes have occurred in recommended treatments over the years.
4	Number of meetings	Year	Multiple meetings for discussing a case have reduced over the years. Or, the MDT meetings have become more efficient over the years.

- *Diagnosis*: the diagnosed cancer influence the treatments recommended to the patient and the number of meetings held for the case.

- The hypotheses for explaining the above hypothetical relations between *diagnosis* and other variables are presented in Table 5-10.

Table 5-10 Hypotheses for the relations between other variables and Diagnosis

Hypothesis	Variable 1	Variable 2	Description
1	Treatment	Diagnosis	A high level treatment, such as surgery, was recommended in situation when the diagnosis of cancer was complex.
2	Number of meetings	Diagnosis	Multiple meetings were held in situation when the diagnosis of cancer was complex.

- *Number of meetings*: the number of meetings held may influence the treatments recommended to the patient, and this is explained by the hypothesis presented in Table 5-11.

Table 5-11 Hypothesis for the relations between Treatment and Number of meetings

Hypothesis	Variable 1	Variable 2	Description
1	Treatment	Number of meetings	A recommendation of treatments for a patient depends on the number of meetings that corresponds for the case.

The structure of the Bayesian network model, we call 'MDT BN', that is determined after deciding on the four relation types between variables is shown in Figure 5-5. The links corresponding to the definitional relations are thicker than the links corresponding to the hypothetical relations.

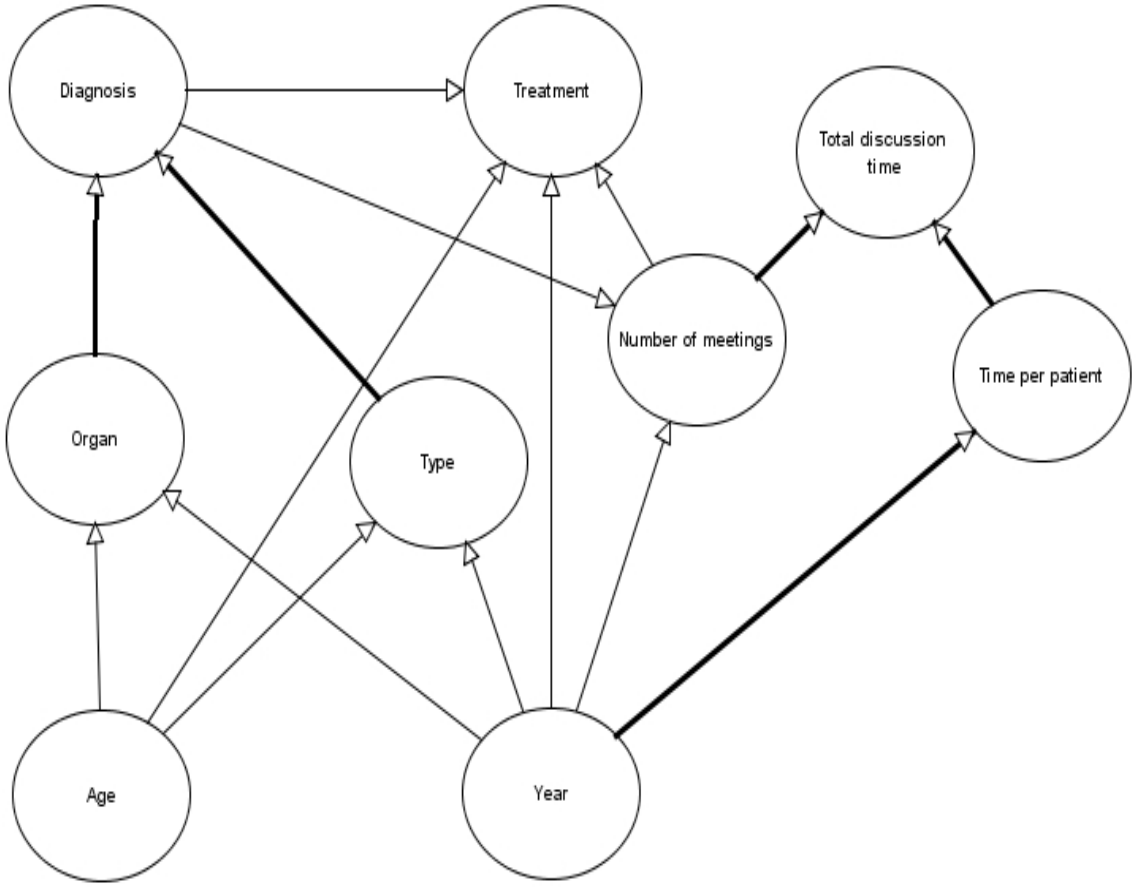


Figure 5-5 Structure of the complete MDT BN model structure from the links of possible relation types

5.5 Summary

In this chapter the structure of a Bayesian network model has been constructed by identifying relations between variables with the help of an expert. The variables represent the key factors in an MDT process which has managed the patients with cancer or suspected cancer treated in the HPB (HepatoPancreaticoBiliary) Centre at the Barts and the London NHS Trust (BLT) between 2005 and 2009. The expert is a clinician who has sufficient experience about the way patients were managed in the MDT meetings held during these years. Since we have only considered knowledge of the expert for constructing the structure, each of the hypothetical links of the model remains uncertain. In the next chapter, we will refine the structure by measuring the

strength of each of the hypothetical relations by taking account of the data collected for the study.

Chapter 6

Establishing the Plausibility of Hypothetical Relations from Data

This chapter focuses on assessing the hypothetical links of an expert-constructed BN model, using data. The MDT BN, introduced in the previous chapter, is used as a case study to present the technique. Each hypothetical link of the model represents a causal relation and is assumed by an expert. However, as many researchers have shown, the use of knowledge often introduces bias. We therefore reassess the existence of these relations using the available data.

The learning method of this study tackles the limitations seen in many existing ones. The method does not rely solely on data nor on expert knowledge, but uses the data to assess the knowledge. Overall, the method follows a Bayesian argument: the set of hypothetical links of the MDT BN determines a set of hypothetical models, including all, some or none of the links. Which of these models explains the data best? To answer this, we build an additional Bayesian network and learn the parameters of the different models, and use these values to calculate the likelihood of the corresponding model.

6.1 Assessing the existence of a relation

Maskery et al. [161] tried to determine if each relation of their BN model, which has initially established using a minimum description length (MDL) scoring algorithm, is important, given the data. The study weighted this importance by (a) deriving a number of BNs from reduced datasets and comparing these BNs with the original BN, and (b) assessing removal cost for links. A link removal cost is defined (i.e. an arc according to

the study) as the change in the MDL score between the original model and the BN in which that particular link has been removed.

In [170] Friedmen et al. applied the Bootstrap method of Efron [171] to different partial structures of a BN, which has been learned from data, for defining confidence about the existence of relations in the BN. The empirical probability of a link, which is defined as the fraction of occurrences in the networks learned from bootstrapped samples, is considered as the strengths of the link. The probability also interprets as the degree of confidence that the link is present in the final network structure that describes the real dependence structure of the original data. A similar technique for assessing relations between variables has also been considered by Heckerman et al. in [172].

In contrast to our work, these earlier studies all assume that a full BN structure is learnt before assessing each single relation, and the methods therefore require multiple structures to be developed for all the variables of the corresponding domain, to found one that is optimal. Further, the work in [170] is yet another example of misuse of null hypothesis testing. For assessing the plausibility of an expert-based association for the available data we use the method introduced in [54], which considers relations between an effect and its causes specified in an expert model as a hypothesis and use data to test this hypothesis against competing hypotheses. The results from the tests reveal the causal relations that are supported by the data.

6.2 The method

The method first identifies different hypotheses in relation to the fragment that has been considered. Each of these hypotheses explains one possible combination of relations for the particular variables.

The aim here is to assign a score to each hypothesis to indicate how well the available data supports that combination.

For simplicity, each hypothesis can be considered as a model, M_i and a test is performed to assign a score based on the available data D. If we denote the model

parameter set by a vector $p_i = (p_{i1}, \ldots., p_{iK})$ of size K, then we can estimate (learn) the probability,$P(p_i|D)$ using Bayes' Theorem:

$$P(p_i|D) = \frac{P(D|p_i) * P(p_i)}{P(D)} \qquad 6.1$$

Now if we consider the data, D, as a vector of observations, i.e. $D = (d_1, \ldots\ldots, d_N)$ of size N, then for the observations Equation 6.1 changes as:

$$P(p_i|d_1, \ldots\ldots, d_N) = \frac{\prod_j P(d_j|p_i) * P(p_i)}{P(d_1, \ldots\ldots, d_N)} \qquad 6.2$$

Suppose that the prior probability for a parameter is always taken as 1, i.e. $P(\boldsymbol{p}_i) = 1$. Then in order to determine how well the parameter set of model M_i explains a number of observations $(d_1, \ldots\ldots, d_N)$, we only compute the likelihood of the model M_i as

$$P(d_1, \ldots\ldots, d_N|p_i) = \prod_j P(d_j|p_i) \qquad 6.3$$

This is what we consider as the score (also known as Bayesian score) for the hypothesis, and the hypothesis which is best with respect to the data D is the one which is assigned with the highest score.

6.2.1 Method advantages

In summary, this method has a number of important advantages:

- It allows the use of expert knowledge to be maximised. Instead of incorporating knowledge partially (e.g. as to determine constrains or to define structural priors over certain relations) a sensible model structure can be specified in order to address all the problems that need reasoning about. The expert will have the flexibility on making assumptions, and this is because there are data available to

check these assumptions.

- Complete data do not have to be processed since the interest here is not learning a structure from all the variables. Only one variable is selected at one time which enables us to prepare less data each time.

- The results from the hypothesis tests can be used to reveal which out of all expert judged relations receive support from data. .

The method is used to check the hypothetical links corresponding to different fragments of the MDT BN model, in which a fragment represents expert's assumed causal relations between an effect and its causes

6.3 Hypothetical links and structure hypotheses

The hypothetical links and corresponding variables can form four sub-models/fragments. Each of these fragments only considers one child and the parents from the MDT BN model. Three of these fragments form from three variables and one from five variables. The four child variables in these fragments are: 1) Number of meetings, 2) Organ, 3) Type, and 4) Treatment.

In this section first we consider each of earlier fragments and revel which of those two relations between the three variables is plausible for the data, and then move into the later fragment.

Hypotheses of three variable fragments

The number of relation combinations possible between a child and two parents is four. We set four hypotheses, each describing one of four relation combinations, to reveal which combination receives support from the available data.

Figure 6-2 depicts the hypotheses regarding the BN fragment shown in Figure 6-1 (for the *Number of meetings* variable). Table 6-1 provides the hypotheses of the variables for the remaining two fragments that consist from three variables.

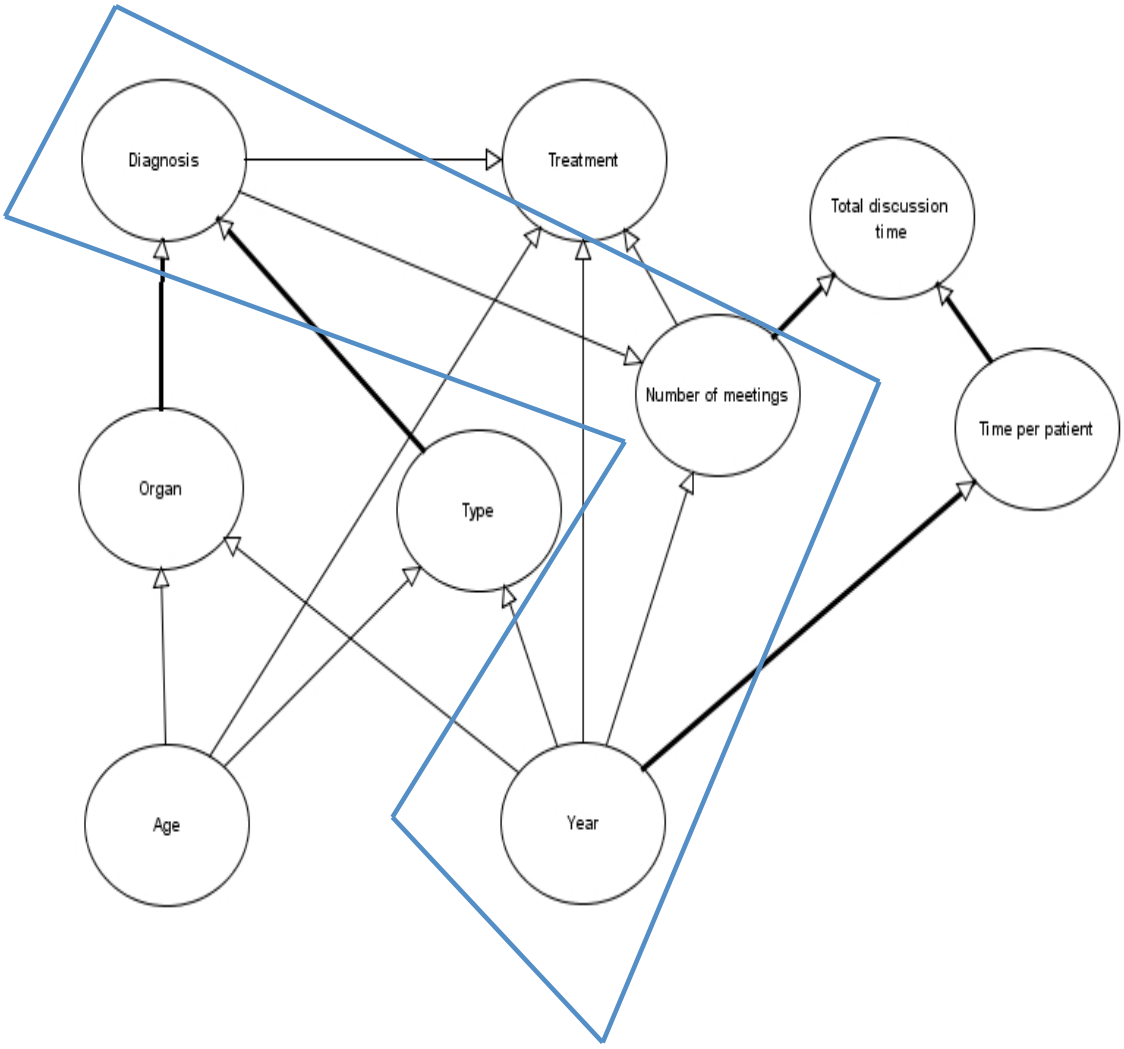
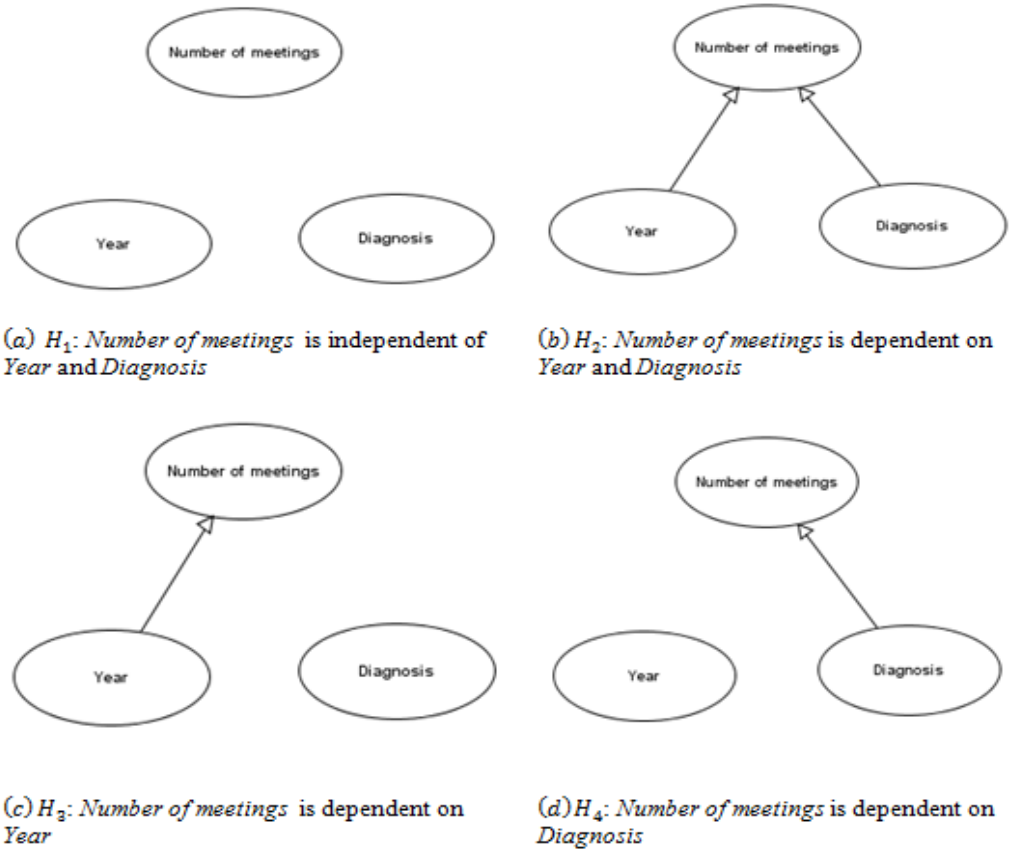


Figure 6-1 BN fragment for learning plausible relation between the Number of meetings and other variables

Table 6-1 Hypotheses of the variables for the fragments relating to Organ and Type

Dependent	Hypothesis	Description
Organ	H_1	Organ is independent of Age and Year.
	H_2	Organ is dependent on both Age and Year.
	H_3	Organ is dependent on Age.
	H_4	Organ is dependent on Year.

Type	H_1	Type is independent Age and Year.
	H_2	Type is dependent on Age and Year.
	H_3	Type is dependent on Age.
	H_4	Type is dependent on Year.



(a) H_1: *Number of meetings* is independent of *Year* and *Diagnosis*

(b) H_2: *Number of meetings* is dependent on *Year* and *Diagnosis*

(c) H_3: *Number of meetings* is dependent on *Year*

(d) H_4: *Number of meetings* is dependent on *Diagnosis*

Figure 6-2 Hypotheses regarding the BN fragment shown in Figure 6-1

Hypotheses of five variable fragments

The number of relation combinations possible for a fragment containing five variables: one child and four parents is sixteen. We set sixteen hypotheses for the Treatment variable, each describing one of sixteen relation combinations, to reveal which combination receives support of existence from the available data:

- the variable *Treatment* is independent of all four parents . *Year, Number of meetings, Diagnosis and Age*. 1 hypothesis.

- the variable *Treatment* is dependent on all four parents. 1 hypothesis.

- the variable *Treatment* is dependent on three out of four parents. 4 hypotheses.

- the variable *Treatment* is dependent on two out of the four parents. 6 hypotheses.

- the variable *Treatment* is dependent on only one parent. 4 hypotheses.

The BN models for the structures stated in the hypotheses (i) and (ii) along with an example BN model for each of the three structures stated from (iii) to (v) are shown in Figure 6-3. The rest of the hypotheses are described in Table 6-2.

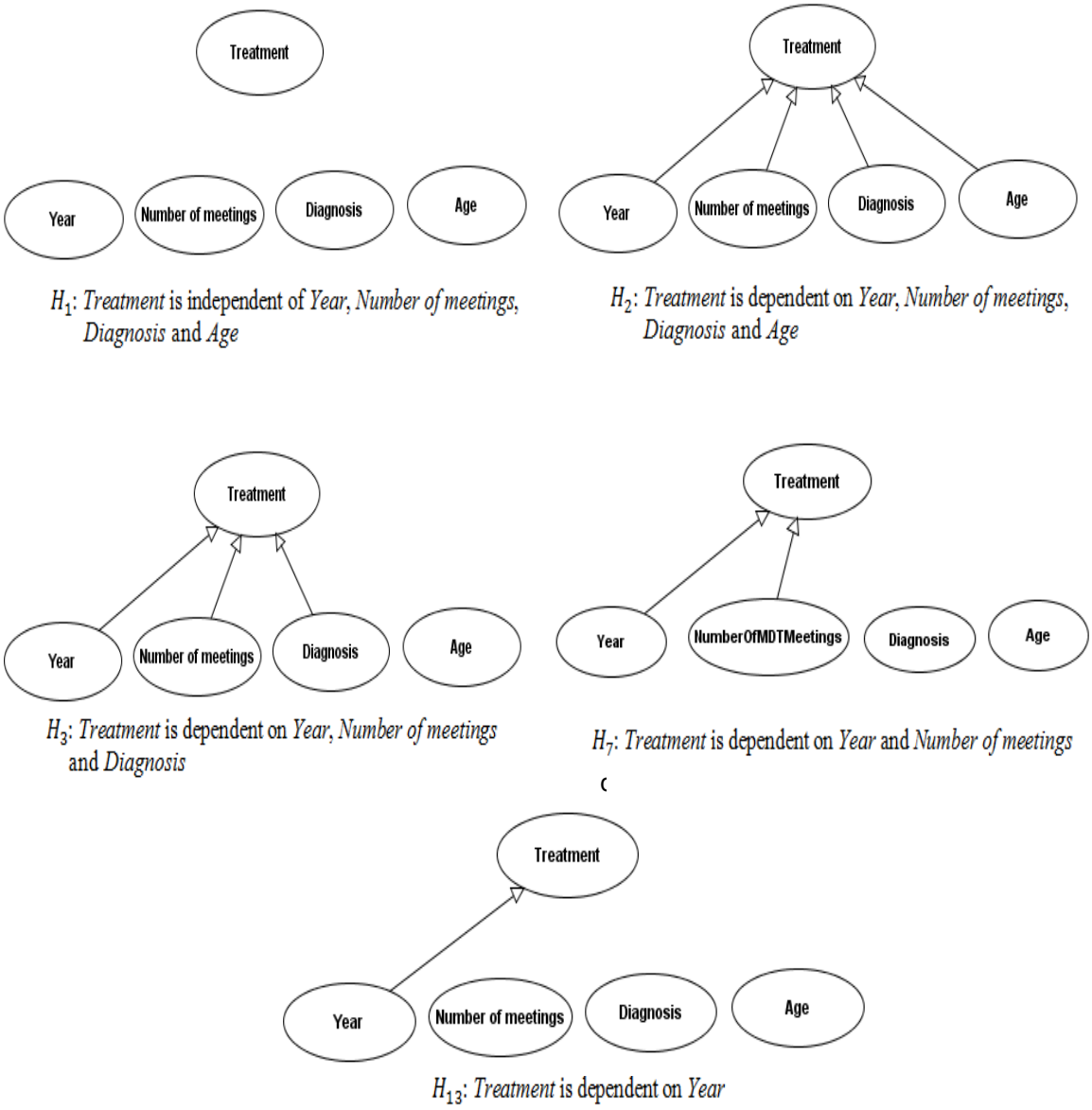


Figure 6-3 Five BN models for representing five different hypotheses: H_1, H_2, H_3, H_7 and H_{13} regarding the *Treatment* fragment

Table 6-2 Hypotheses that do not appear in Figure 6-3

Hypothesis:	Treatment is dependent on
H_4	Year, Age and Number of meetings
H_5	Age , Number of meetings and Diagnosis
H_6	Year, Age and Diagnosis
H_8	Year and Diagnosis
H_9	Age and Number of meetings
H_{10}	Year and Age
H_{11}	Age and Diagnosis
H_{12}	Number of meetings and Diagnosis
H_{14}	Age
H_{15}	Number of meetings
H_{16}	Diagnosis

6.4 Parameters and explanation of data

First the parameters of a possible structure – corresponding to a hypothesis – are learned using data in order to determine how well the structure explains the data. Each parameter is the conditional probability for each outcome of the child for each combination of values of its parents. Figure 6-4 shows the structure of an example of the additional Bayesian networks constructed to learn a conditional probability parameter. One of these additional networks is needed for each parameter of the model; together the additional networks are used for determining how well the parameters have explained the data.

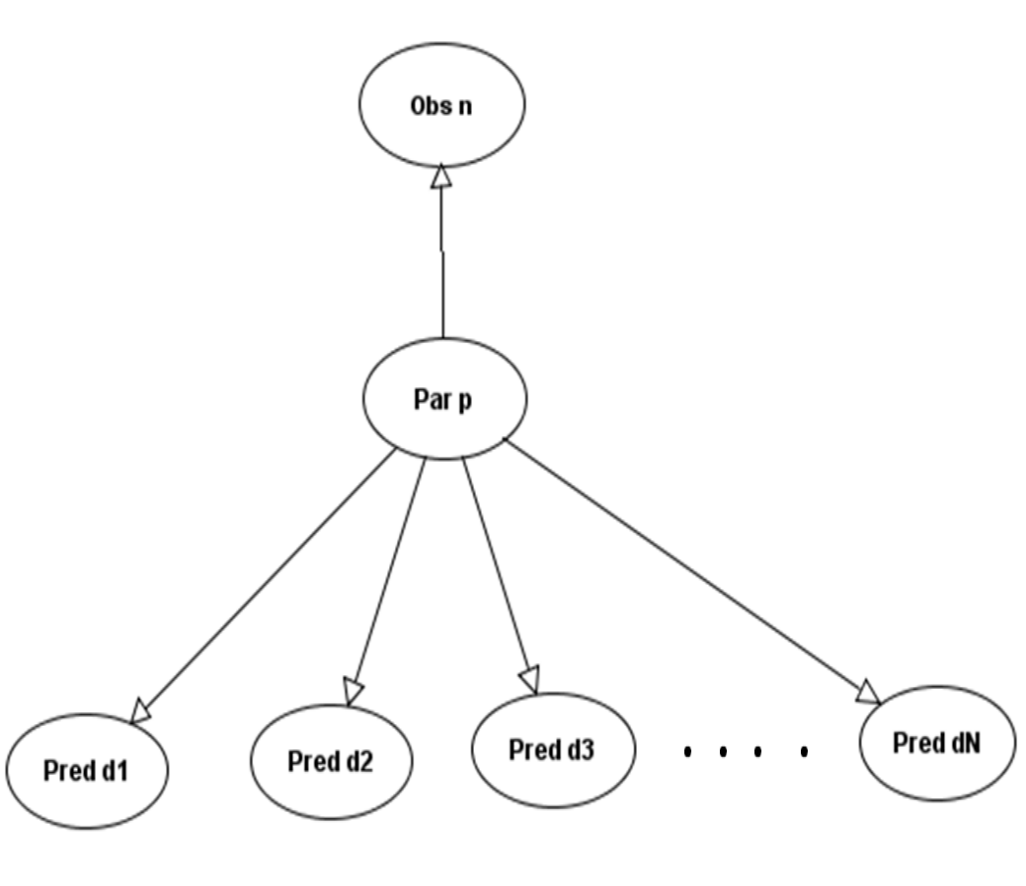


Figure 6-4 A Bayesian network model for parameter estimation

The purpose of each of the nodes in this Bayesian network is explained as follows:

- Estimating parameter using Equation 6.2: the $Par\ p$ node is for estimating the parameter. The prior for the parameter is a $\sim Uniform\ (0,1)$, and the posterior probability computes after entering observation into the network.

- Entering observation: the $Obs\ n$ node is the observation node for $Par\ p$. The node is $\sim Binomial(n, p)$. For example, in the case of *Year* → *Number of meetings*, n is the number of meetings held during the particular year and p is the probability of the meeting category.

- Calculating the probability of data: once the posterior parameter is estimated, then we calculate the probability of observing each data point conditional on that the model parameter learned. This is done by creating a node for each data point

and entering fixed bins for $[< d][d][> d]$, where d is the data point. These prediction nodes for each data point are denoted as $Pred\ d_i$, where $i = 1, \dots, N$ and, N is the number of data points.

6.5 Selection of plausible relations – three variables

This section provides an example of the calculation using the fragment for the *Number of meetings* variable, which may depend on the variables *Diagnosis* and *Year*.

For each possible model of the variables considered, the posterior probability of each of the prediction nodes $Pred\ d_i$of the parameter models (i.e. the additional BNs constructed for learning parameters) are used to calculate the likelihood of all data given the model using Equation 6.3. The normalised score for each model indicates if the model is best for explaining the data.

Table 6-3 presents the distribution of meeting categories for patients of each year of discussion and diagnosis combination. In particular, the dataset is for the variables of the fragment *Number of meetings*. We use the dataset to check each of the models presented in Figure 6.2. The model for the first hypothesis, H_1: *Number of meetings* is independent of *Diagnosis* and *Year*, corresponds to four parameters, of which three parameters are learned. This is because, the parameters for three out of the four meeting categories ensure sufficient data points to understand the model's overall fit to the data.

There are forty data points per meeting category (see Table 6.3), and thus, each parameter model has forty prediction nodes. In most cases, a prediction node has three fixed bins (see section 6.4), but when a recorded value is '*0*' the node for this is set with bins $[0][1 - 11][12 - infinity]$ to calculate the probability of the value. Table 6-4 presents the probabilities of observing the data points conditional on that the model parameters are learned.

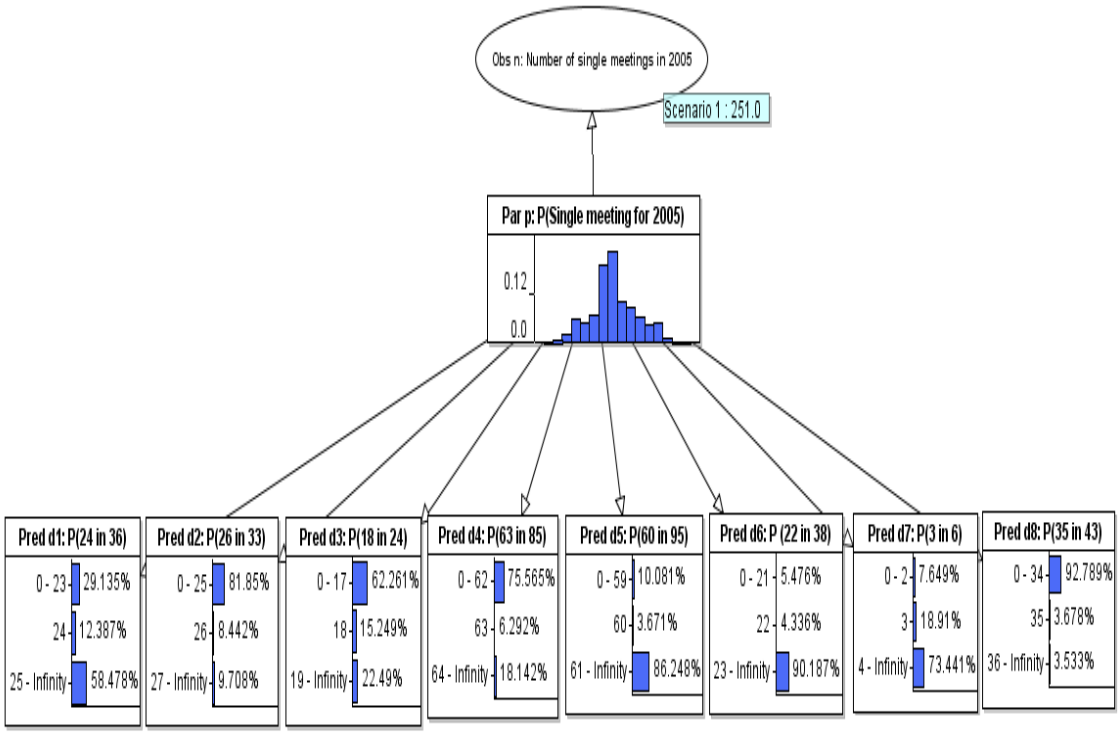
The model for the second hypothesis, H_2: *Number of meetings* is dependent on *Diagnosis* and *Year,* corresponds to (40×4) = 160 parameters, of which 120 parameters are learned. Each parameter model constructed for learning has only one data point. Figure 6-5 (c) illustrates the parameter model for learning the probability of a single meeting given that the discussion year is 2005 and the diagnosed cancer is benign pancreas. Appendix A.1 provides the posterior probability of the data point of this model along with the posterior probabilities of the data points of other parameter models constructed to assess this hypothesis.

Further, the models for H_3: *Number of meetings* is dependent on *Year* and H_4: *Number of meetings* is dependent on *Diagnosis* correspond to 20 parameters and 32 parameters, respectively. Similar to the above, based on any combination of three meeting categories, the numbers of parameters select for learning H_3 is 15 and for learning H_4 is 24. Figure 6-5 illustrates two of the parameter models constructed for these hypotheses. In particular, Figure 6-5 (a) illustrates that an observation regarding the year of discussion provides updated probability for receiving single meetings, which then remains constant over all data points to calculate the likelihood of each. Figure 6-5 (b) illustrates that an observation regarding the diagnosed cancer provides updated probability for single meetings which then remains constant over all data points to compute the posterior probability of each. The posterior probabilities of the data points from the all the parameter models of H_3 and H_4 can be found in Appendix A.1.

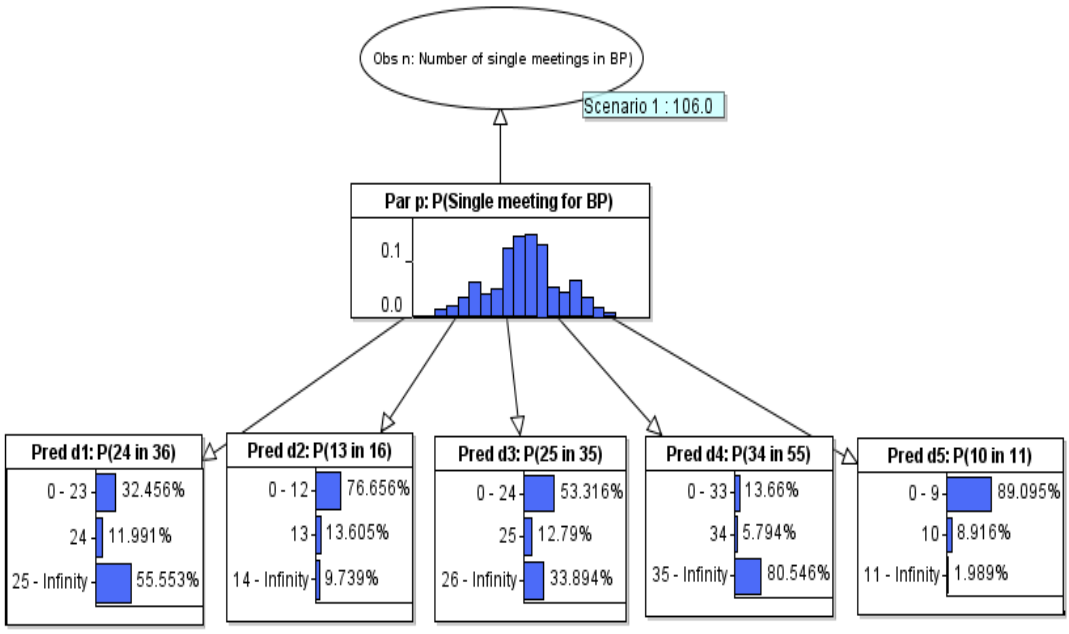
Table 6-3 Number of meetings and total data per meeting category for patients of each Year and Diagnosis combination

Year	Diagnosis	Number of meetings				Total		Year	Diagnosis	Number of meetings				Total
		1	2	3	4 or more					1	2	3	4 or more	
2005	BP	24	12	0	0	36		2007	ML	87	39	5	3	134
	BL	26	7	0	0	33			MGB	29	14	1	0	44
	BGB	18	6	0	0	24			Multiple	9	1	1	0	11
	MP	63	21	1	0	85			Unknown	23	8	2	0	33
	ML	60	31	4	0	95		2008	BP	34	19	2	0	55
	MGB	22	15	1	0	38			BL	29	11	0	0	40
	Multiple	3	3	0	0	6			BGB	10	9	1	0	20
	Unknown	35	8	0	0	43			MP	61	24	2	0	87
2006	BP	13	3	0	0	16			ML	61	40	4	0	105
	BL	13	1	0	1	15			MGB	27	19	2	0	48
	BGB	7	6	0	0	13			Multiple	11	5	0	0	16
	MP	26	7	2	0	35			Unknown	46	23	3	1	73
	ML	44	10	3	0	57		2009	BP	10	0	1	0	11
	MGB	15	3	0	0	18			BL	3	2	1	0	6
	Multiple	3	0	0	0	3			BGB	8	2	0	0	10
	Unknown	19	2	0	0	21			MP	11	5	0	0	16
2007	BP	25	10	0	0	35			ML	15	16	0	2	33
	BL	14	4	1	0	19			MGB	1	4	1	0	6
	BGB	19	4	0	0	23			Multiple	1	0	0	0	1
	MP	64	23	3	1	91			Unknown	15	6	0	0	21

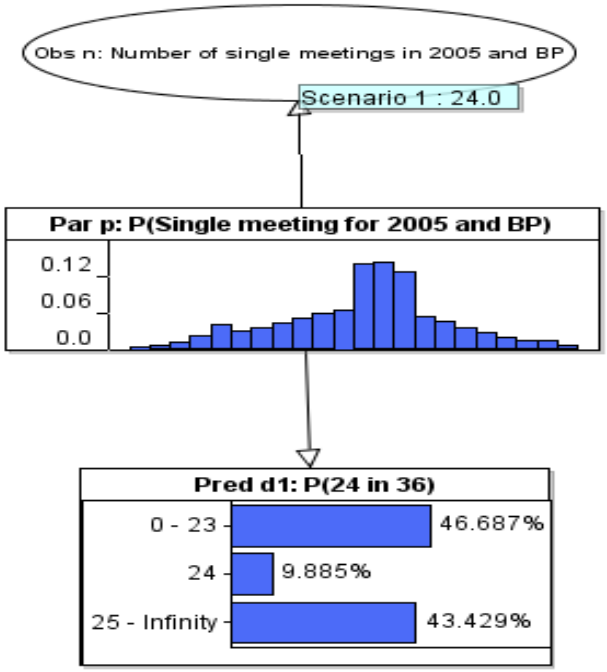
The normalised Bayesian score for each possible model for the variables *Number of meetings*, *Year* and *Diagnosis* is presented in Table 6-5. Results are obtained by calculating the joint probability of data using Equation 6.3 and normalising the values that are found for all competing models. As can be seen from the table, Bayesian score of the model that represents H_2 is higher than the other models, i.e. the model is the best for the data observed.



(a) A model constructed for learning a parameter for H3



(b) A model constructed for learning a parameter for H4



(c) A model constructed for learning a parameter for H2

Figure 6-5 Bayesian parameter learning networks for determining the best structure for the variables Number of meeting, Year and Diagnosis

Table 6-4 Probabilities of data points obtained per learned parameter for H_1 of the Number of meetings fragment

Year	Diagnosis	P(data point | Number of meetings =			Year	Diagnosis	P(data point| Number of meetings =		
		1	2	3			1	2	3
2005	BP	0.135	0.115	0.359	2007	ML	0.053	0.072	0.144
	BL	0.065	0.104	0.391		MGB	0.118	0.114	0.357
	BGB	0.138	0.169	0.504		Multiple	0.174	0.108	0.233
	MP	0.047	0.070	0.217		Unknown	0.144	0.135	0.173
	ML	0.051	0.060	0.146	2008	BP	0.068	0.072	0.254
	MGB	0.056	0.048	0.367		BL	0.114	0.136	0.321
	Multiple	0.205	0.171	0.841		BGB	0.044	0.054	0.325
	Unknown	0.023	0.049	0.295		MP	0.082	0.090	0.254
2006	BP	0.122	0.163	0.632		ML	0.009	0.011	0.163
	BL	0.072	0.038	0.650		MGB	0.028	0.033	0.237
	BGB	0.123	0.090	0.689		Multiple	0.208	0.204	0.632
	MP	0.109	0.084	0.183		Unknown	0.063	0.085	0.185
	ML	0.039	0.022	0.140	2009	BP	0.075	0.025	0.233
	MGB	0.082	0.121	0.597		BL	0.205	0.318	0.147
	Multiple	0.313	0.363	0.917		BGB	0.209	0.247	0.750
	Unknown	0.015	0.029	0.548		MP	0.208	0.204	0.632
2007	BP	0.132	0.146	0.369		ML	0.004	0.008	0.391

	BL	0.175	0.164	0.318			MGB	0.015	0.052	0.147
	BGB	0.062	0.098	0.518			Multiple	0.677	0.713	0.972
	MP	0.079	0.072	0.213			Unknown	0.177	0.188	0.548

The method is applied in the same way to each of the remaining two fragments of three variables to select the best models among those that are possible; under each possible structure, a BN model constructs for each considered parameter. After estimating the parameter from data, the posterior probability of each corresponding data point is computed and finally, a Bayesian score from the joint probability of data is used to determine the best combination of relations for the variables considered. More calculation details regarding the hypotheses of the variables appeared in these fragments can be found in Appendix A.2 and in Appendix A.3.

The normalised Bayesian score for each possible model for the variables of the above three fragments is presented in Table 6-5. As can be seen from the table, the learning method generates a higher score for H_2 under all three fragments. Since a higher score means stronger evidence in supporting the corresponding structure for the considered data, we can conclude that each of the three variables: *Number of meetings*, *Organ* and *Type* is dependent on both the parents *Age* and *Year*.

Table 6-5 Joint probability and score per hypothesis per fragment

Outcome	**Number of meetings**		**Organ**		**Type**	
Hypothesis	Joint probability	Score	Joint probability	Score	Joint probability	Score
$\boldsymbol{H_1}$	1.2829E-105	5.47388E-16	1.2236E-101	7.08647E-12	6.7281E-123	1.25859E-52
$\boldsymbol{H_2}$	2.34372E-90	0.999999953	1.7267E-90	0.99999655	5.34574E-71	0.999999868
$\boldsymbol{H_3}$	1.10872E-97	4.73059E-08	5.89354E-96	3.41316E-06	7.05914E-78	1.32052E-07
$\boldsymbol{H_4}$	2.4982E-100	1.06591E-10	6.32575E-98	3.66347E-08	5.5023E-114	1.02928E-43

6.6 Selection of plausible relations – five variables

The fragment of the MDT BN model for five variables is shown in Figure 6-6. With four hypothetical links, the expert has shown that *Treatment* depends on *Age*, *Year*, *Diagnosis* and *Number of meetings*. The data for learning plausible relations are the number of patients for each treatment option under each combination of values of all four parents. There are sixteen competing hypotheses (as explained in section 6.3). Similar to the previous section, we make use of the data to learn the parameters for each of these sixteen models.

Since the number of parameters required for H_1: *Treatment* is independent of *Year, Number of meetings, Diagnosis* and *Age*, is seven, as above, six of them are only needed to learn in order to assess the hypothesis. An additional BN network is constructed for learning each parameter for the hypothesis H_1; since this hypothesis has the fewest parameters it associates with more data points than any other additional BN constructed for learning a parameter for any other hypothesis. In general, there is a decrease in the number of data points related to a parameter which corresponds to a hypothesis that corresponds to a greater number of parents. Each of the 6720 parameters (i.e. a conditional probability of a state for one of the 1120 possible combinations of parents) for H_2: *Treatment* is dependent on *Year, Number of meetings, Diagnosis* and *Age*, therefore, had only one data point in association.

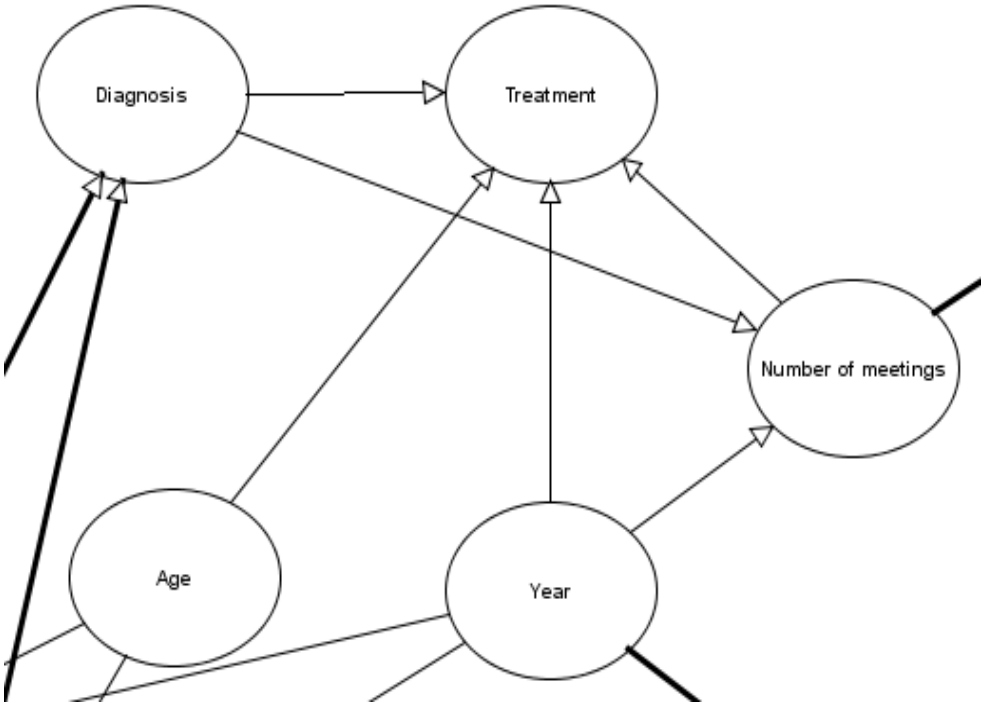


Figure 6-6 Structure of the MDT BN model fragment for learning plausible relations between five model variables

Unlike the case of learning a parameter for H_2, where an observation to a BN relates to only one data point so the posterior probability of one data point derives at once, for learning a parameter for H_1, an observation relates to 1120 data points. This creates complications in calculating the probabilities given the observation; the time and memory needed to calculate the posterior probabilities for all the data points is unattainable.

Java and AgenaRisk API, Java application program interface, are applied in this study to (a) construct the parameter learning BNs, (b) enter observations into the BN from a data file, (c) propagate the BN, and (d) generate the posterior probability of each data point (e) save the probabilities in a data file. However, when applied to the no causes case, H_1, for a five variable fragment we are unable to complete learning any parameter for H_1; the allocated memory runs out before the process could complete.

To get around this problem I have deleted those combinations of parents for which the total data (i.e. the number of patients observed for all treatment options) is found to be zero, and learned the parameters using the remaining data. This makes no difference in learning and does not require making any approximation. In particular, 960 combinations of parents have been deleted in total.

The joint probabilities of data calculated for each of the sixteen hypotheses are very small value (very close to '0'). We therefore applied logarithmic ($log10$) transformations to the probabilities to bring the values closer to numbers that can be used to compute a normalised score for each hypothesis. In particular, four steps are followed to calculate a score for a hypothesis:

- The joint probability of data for each hypothesis is transformed by applying $log10$.

- The $log10$ Bayes Factor (*BF*) is calculated using Equation 2.3 for the hypotheses. In each case, the hypothesis that is chosen is H_3, as it has the highest log10 BF among all calculated values.

- We then remove $log10$ by applying $antilog10$ to each calculated *BF*.

- Finally, the BFs are normalised so that the sum of the BFs is equal to '1'.

Table 6-6 summarises the scores of the sixteen hypotheses for the variables *Treatment*, *Year, Number of meetings, Diagnosis* and *Age*. Clearly, the method produces a higher score for H_3 than any other hypothesis, which shows a higher support about the structure from data. So considering the scores in Table 6-6, we obtain the highest evidence to conclude that *Treatment* is dependent on three parents: *Year*, *Number of meetings*, *Diagnosis*, but not four that the expert has thought.

Table 6-6 Normalised scores of the sixteen competing hypotheses one of which assumes to be the best in terms of the data

Hypothesis	Likelihood (in log10)	BF	BF (revised)	Score
H_1	-884.368	-222.263	5.455E-223	4.9016E-223
H_2	-733.794	-71.689	2.04752E-72	1.8398E-72
$\boldsymbol{H_3}$	-662.105	0	1	0.898552375
H_4	-788.010	-125.905	1.244E-126	1.1178E-126
H_5	-673.525	-11.420	3.80258E-12	3.41682E-12
H_6	-705.348	-43.243	5.7107E-44	5.13136E-44
H_7	-848.134	-186.029	9.3535E-187	8.4046E-187
H_8	-665.496	-3.391	0.000406231	0.00036502
H_9	-832.845	-170.741	1.8171E-171	1.6328E-171
H_{10}	-812.405	-150.300	5.0077E-151	4.4996E-151
H_{11}	-668.310	-6.205	6.2341E-07	5.60167E-07
H_{12}	-663.054	-0.949	0.112494328	0.101082045
H_{13}	-867.836	-205.731	1.8589E-206	1.6704E-206
H_{14}	-851.743	-189.638	2.3015E-190	2.068E-190
H_{15}	-872.490	-210.385	4.1199E-211	3.7019E-211
H_{16}	-676.489	-14.384	4.12934E-15	3.71043E-15

6.7 Revised model structure

The revised structure of the MDT BN model, which we constructed in Chapter 5 and presented in Figure 5-5, is shown in Figure 6-7. The definitional links, those presented with thicker links, of the original model represent deterministic relations between the model variables, so no further revision was made for these links. The learning method is applied on to four fragments of the model in each of which the variables remain connected with hypothetical links. In one of these fragments the revised structure turns out to be dissimilar than which the expert has initially assumed; according to the expert's assumptions, *Age* is one of the four factors to have strong influence on *Treatment* but learning from data does not provide enough evidence in support of a structure that considers this relation. In each of the remaining three fragments, the expert has assumed relations between variables appropriately.

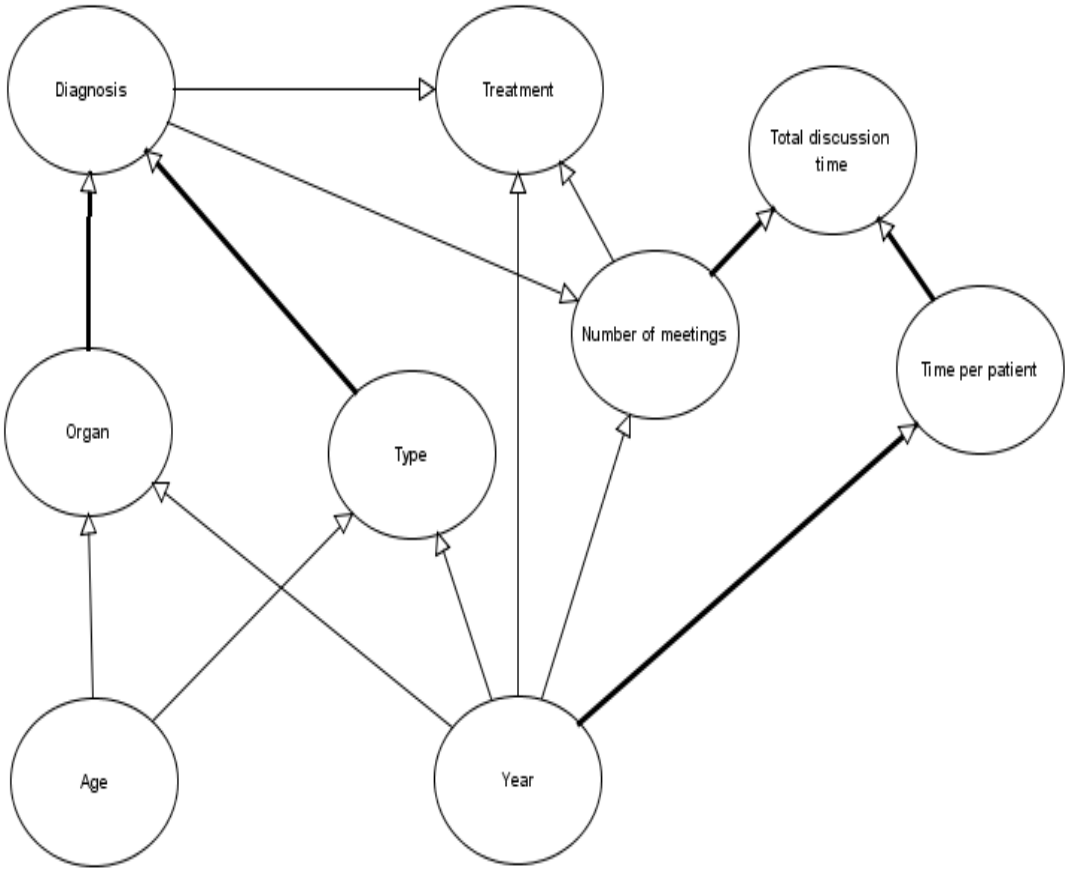


Figure 6-7 Revised structure of the MDT BN after learning the data supported relations.

6.8 Summary

This chapter demonstrated how relationships between variables in an expert constructed Bayesian network model can be checked to establish their plausibility with respect to data. The method applied enabled us to derive the best structure for each effect variable of the expert model in terms of data. These established structures were then combined to create a full model structure. The results from the hypothesis tests reveal that the clinician was right about most of the links. In particular, the revised MDT BN model has one less link than those appeared in the expert model. The next chapter focuses on conditional dependencies among variables and assesses the quality of each link of the BN.

Chapter 7

Evaluating the Strength of Associations

Following the steps described in Chapters 5 and 6, the MDT model is both causally coherent – with support from a domain expert – and statistically valid – with support from data. As a result, the MDT model provides a framework to analyse the MDT data. The aim of this chapter is to analyse and determine the strength each association, and show how to form and assess hypotheses of clinical interest.

7.1 Impact of MDT meetings

As described in Chapter 5, the MDT BN model represents an MDT process through which recommendations were provided to cancer patients in the HPB (HepatoPancreaticoBiliary) centre at the Barts and the London NHS Trust (BLT). We wish to use the model to gain information against the success of MDT meetings for making treatment recommendations. For example, we are interested to know whether the MDT meetings have become more efficient over time.

The increased interest in a coordinated plan for the management of cancer patients has made MDT meetings popular in oncology. Many researchers have shown that the management of cancer patients through discussions in MDT meetings improves outcomes.

Gabel et al. [168] demonstrated that MDTs have improved patients' satisfactions by decreasing the time from diagnosis to treatment. Burton et al. [173] have concluded that MDT discussion of staging with magnetic resonance imaging and following implementation of preoperative treatment have significantly reduced the rate of positive circumferential resection margins.

For finding evidence of the benefit of a lung cancer MDT meeting, the authors in [174] have used Chi-square tests and *t* tests along with a logistic regression model. The authors in [154] have also used Chi-square test while analysing the decisions of MDT meetings to skin cancer patients in order to find evidence of benefit. Both studies have considered a P value < 0.05 to be statistically significant.

We discussed the limitations of these measures (P-values and Confidence Intervals) of frequentist statistics in detail in Chapter 2, and instead propose a Bayesian approach to evaluate the impact of the MDT meetings. Specifically, we assess the strength of each relation of the MDT BN model. Overall, for each relation, firstly we frame hypotheses about the effect, secondly learn the conditional probabilities and compute Bayes Factors, and finally, analyse these values. Provided that appropriate hypotheses are chosen, this analysis can show the changes that have occurred in the operation of MDT for making treatment recommendations and whether these changes have improved the efficiency of the MDT process over the years covered by the data.

7.2 Conditional probabilities from data

In this section, we explain construction of the Bayesian network that is used (in Section 7.4) to compute the Bayes Factors of hypotheses about the strength of relations. These relations are determined by the conditional probabilities that are the parameters of the BN MDT model. Each conditional probability is the probability of a state of a child variable given the state of its parents. The probabilities are calculated from data. For example, in order to compute the probability of a cancerous organ given the age group based on the link $Age \rightarrow Organ$, we count the total number of cancerous organs for patients in the particular age group as shown in Table 7-1. Each value included in the

row 'Total' includes the total number of cancerous organs for patients in each of the seven age groups during the period of the study.

Table 7-1 Counts of the cancerous organ for patients in the corresponding age group

Organ	Age						
	Under 46	46 – 54	54 – 60	60 – 66	66 – 71	71 – 77	77+
Pancreas	74	92	85	87	84	73	81
Liver	107	91	82	98	93	102	85
Bile duct	21	21	24	34	35	40	35
Gallbladder	9	2	11	14	15	16	15
Total	211	206	202	233	227	231	216

For each of the age groups in Table 7-1, there are four parameters. However, instead of just calculating the percentage of the patients in each category, we make use of multinomial distribution by constructing a multinomial BN model (Figure 7.1). The multinomial distribution is a generalisation of the binomial distribution. The binomial distribution gives the probability of each outcome in n independent trials of a two outcome process whereas the multinomial distribution gives the probability of each outcome in n independent trials of a process in which there are more than two outcomes.

In Figure 7-1, each parameter of the model is depicted as a node and denoted as $Par\ p_i$, where $i = 1, \ldots, N$ and N is the number of parameters. For the purpose of this study, we consider the prior for each parameter as uniform. The posterior distribution of parameters is computed after entering observations into the model. The observation node for each parameter is denoted as $Obs\ d_i$, where $i = 1, \ldots, (N-1)$ and N is the number of data points. We modelled the observation node for $Par\ p_1$ using the expression in Equation 7.1 and the observation nodes for other parameters using the expression in Equation 7.2. With these expressions the model assumes that the observed data for each parameter is taken as Binomial, which is known to be an appropriate underlying sampling distribution when the number of cases and the size of the population are available [176].

$$Binomial\,(n1, p_1) \tag{7.1}$$

$$Binomial\left(n1 - \sum_{i=1}^{j-1} n_i, \frac{p_j}{\sum_{i=j}^{N} p_i}\right) \tag{7.2}$$

where n_1 is the total data for the age group and n_i is the remaining total data at position i, and $j = 2, \ldots\ldots, N-1$. Each node for total data is denoted as $Tot\ n_i$, where $i = 1, \ldots., N$ and N is the number of data points.

Having seen the observed data, the priors of the parameters update with a posterior distribution. A constraint is then used to set the sum of the probability over all the values equal to 1 (i.e. $\sum_{i=1}^{4} p_i = 1$).

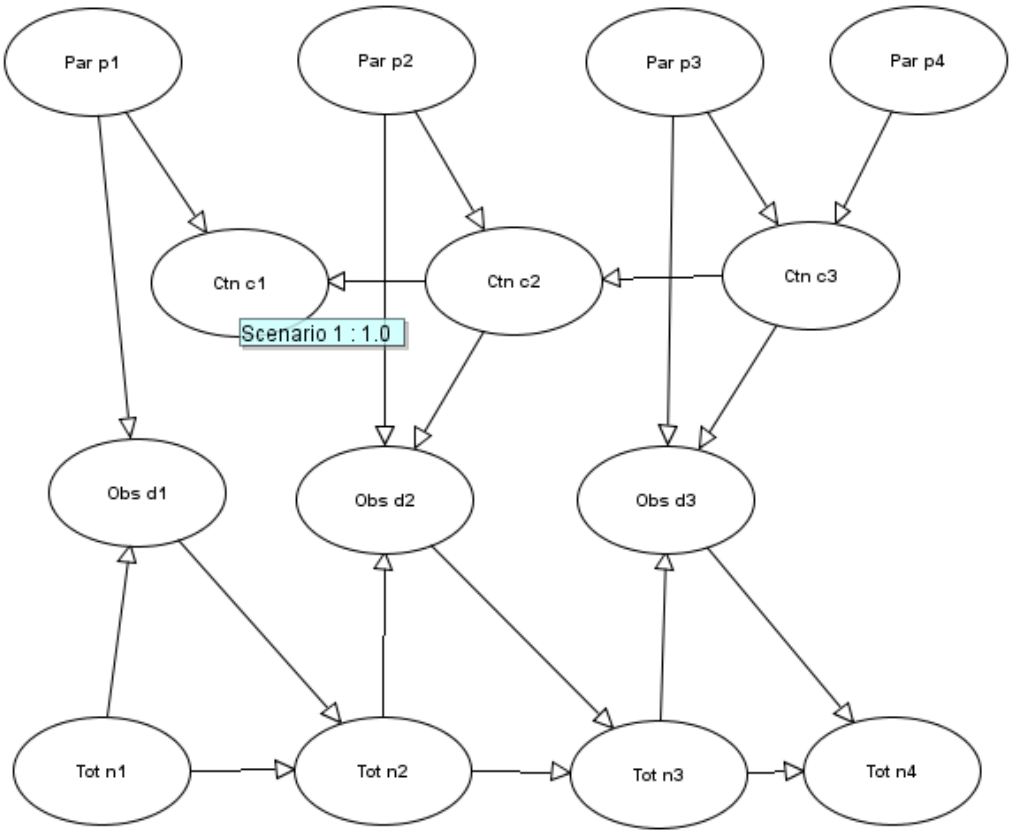


Figure 7-1 A multinomial BN model for estimating parameters from data

Figure 7-2 shows the underlying expressions used in the nodes of the multinomial model constructed to compute the posterior distribution of four parameters.

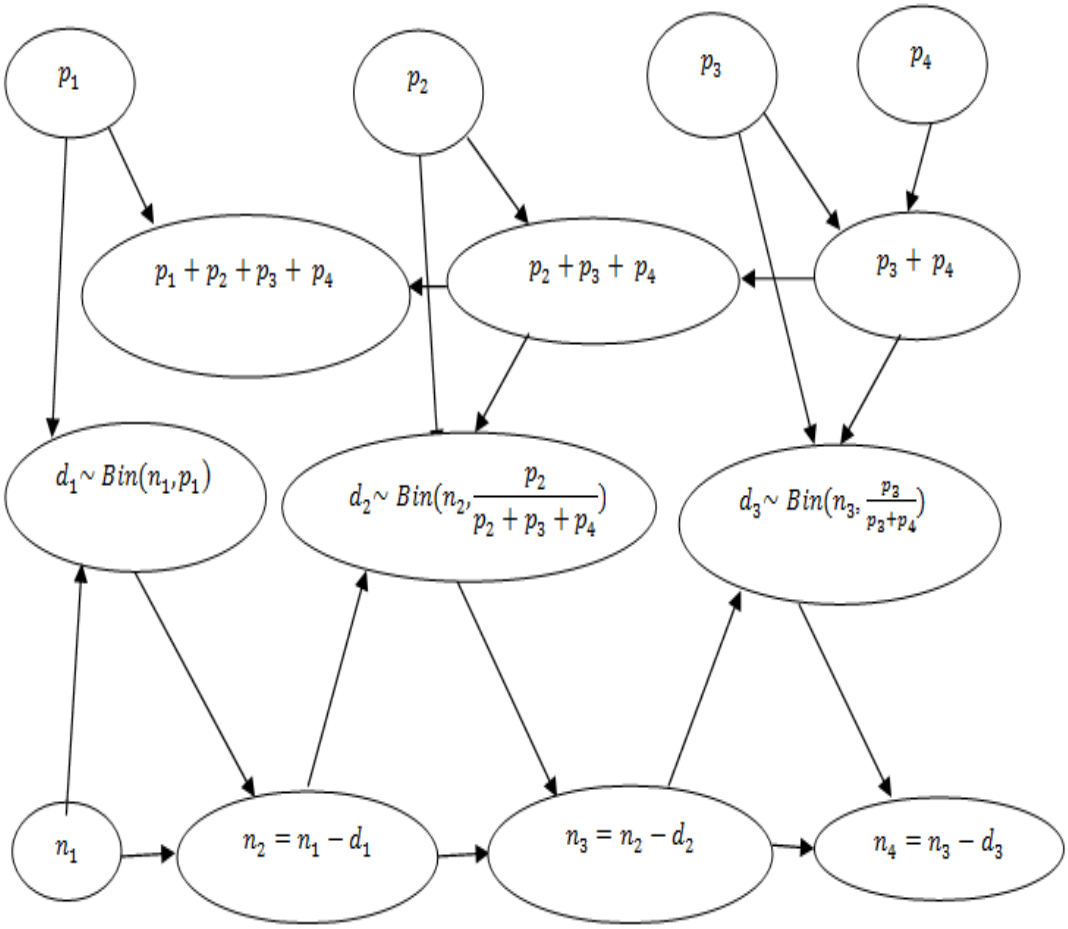


Figure 7-2 Underlying expressions used for nodes in a multinomial BN model

Table 7-2 provides the posterior probability (i.e. the mean of each distribution) for a cancerous organ given the age group of a patient. The complete conditional probability table for *Organ* is obtained from seven four-parameter multinomial BN models.

Table 7-2 Probability of each cancerous organ given the age group

Organ	Age						
	Under 46	46 – 54	>54 – 60	>60 – 66	>66 – 71	>71 – 77	77+
Pancreas	0.348	0.441	0.417	0.371	0.368	0.316	0.373
Liver	0.502	0.438	0.403	0.418	0.407	0.438	0.39
Bile duct	0.106	0.107	0.121	0.148	0.156	0.174	0.164
Gallbladder	0.044	0.014	0.058	0.063	0.069	0.072	0.073

The remaining relations in the BN model correspond to the variables: (a) *Age,* (b) *Year* and (c) *Diagnosis.* Tables 7-3 to 7-5 provide calculated probabilities (posterior means) for the states of each variable that depends on one of these three variables in a corresponding link.

Table 7-3 Probability of each cancer type given the age group

From	To	States	Means						
			Under 46	46 – 54	>54 – 60	>60 – 66	>66 – 71	>71 – 77	77+
Age	Type	Benign	0.64	0.51	0.36	0.31	0.28	0.23	0.3
		Malignant	0.34	0.46	0.59	0.66	0.69	0.73	0.65
		Unknown	0.02	0.03	0.05	0.03	0.03	0.04	0.05

Table 7-4 Probabilities over the states of a) Organ, b) Type, c) Number of meetings and d) Treatment given year

From	To	States	Means				
			2005	2006	2007	2008	2009
Year	Organ	Pancreas	0.37	0.34	0.36	0.43	0.29
		Liver	0.43	0.46	0.45	0.37	0.52
		Bile duct	0.14	0.14	0.14	0.15	0.09
		Gallbladder	0.06	0.06	0.05	0.05	0.1
	Type	Benign	0.37	0.37	0.31	0.41	0.42
		Malignant	0.61	0.61	0.67	0.53	0.53
		Unknown	0.02	0.02	0.02	0.06	0.05
	Number of meetings	1	0.69	0.77	0.69	0.62	0.6
		2	0.28	0.18	0.26	0.33	0.34
		3	0.02	0.03	0.04	0.04	0.04
		4 or more	0.01	0.02	0.01	0.01	0.02
	Treatment	Chemotherapy	0.06	0.03	0.09	0.06	0.08
		Combination	0.21	0.16	0.11	0.09	0.05
		None	0.27	0.24	0.22	0.29	0.29
		Palliative	0.21	0.27	0.25	0.19	0.23
		Surgery	0.14	0.17	0.21	0.21	0.15
		Intervention radiology	0.04	0.04	0.03	0.04	0.02
		Watchful waiting	0.07	0.09	0.09	0.11	0.18

Table 7-5 Probability of each meeting category given Diagnosis

From	To	States	Means							
Diagnosis			BP	BL	BGB	MP	ML	MGB	Multiple	Unknown
	Number of meetings	1	0.67	0.73	0.67	0.71	0.62	0.6	0.68	0.71
		2	0.29	0.22	0.3	0.25	0.32	0.35	0.25	0.25
		3	0.03	0.03	0.01	0.03	0.04	0.04	0.05	0.03
		4 or More	0.01	0.02	0.02	0.01	0.02	0.01	0.02	0.01

7.3 Impossible combination and learning conditional probabilities

This section describes a variation of the method of Section 7.2 to allow for values that are known to be zero by definition (c.f. just observed to be zero). An example arises in the link *Diagnosis* → *Treatment* of the revised MDT BN, where some counts are always zero. For example, when a patient suffers with a benign cancer the MDT meetings never recommend chemotherapy as a treatment; the combination and palliative treatment options are also impossible. Table 7-6 states the number of patients with the particular treatment option for each particular diagnosis. An empty cell of the table indicates for that for the particular diagnosis no such treatment option was possible. Since such parameters are defined to be zero, we do not want to infer a small non-zero value. Moreover, by omitting these parameters and by computing a probability only when a treatment recommendation is a possible option, we increase the efficiency of the method.

Table 7-6 Counts of the recommended treatments for patients in the corresponding diagnosis option

From	To	States	Means							
			BP	BL	BGB	MP	ML	MGB	Multiple	Unknown
Diagnosis	Treatment	Chemotherapy	-	-	-	26	55	9	-	1
		Combination	-	-	-	70	82	36	-	4
		None	53	68	21	38	60	11	5	127
		Palliative	-	-	-	107	150	59	2	10
		Surgery	23	15	37	68	57	37	14	12
		Intervention radiology.	10	9	-	2	20	2	-	2
		Watchful waiting	65	20	16	3	0		5	35

The steps required are:

- Consider the diagnostic options one at a time and determine how many parameters are there to learn;
- Construct a multinomial model for learning the parameters. The model corresponds to number of the parameters that we need to learn.
- Enter evidence into the models, and calculate the posterior probability distribution over each parameter given evidence using Bayes theorem.
- Return the mean value of the distribution of each parameter as its calculated posterior probability.

Figures 7-3(a) and 7-3(b) depict the multinomial models constructed for learning parameters in relation to the diagnoses: benign pancreas (BP) and malignant pancreas (MP), respectively. The number of parameters of 7-3(a) is not same as the number of parameters of 7-3(b). In particular, for $Diagnosis = BP$, the probability is only computed for four possible treatment options, whereas, for $Diagnosis = MP$ we can compute the probability for every treatment option.

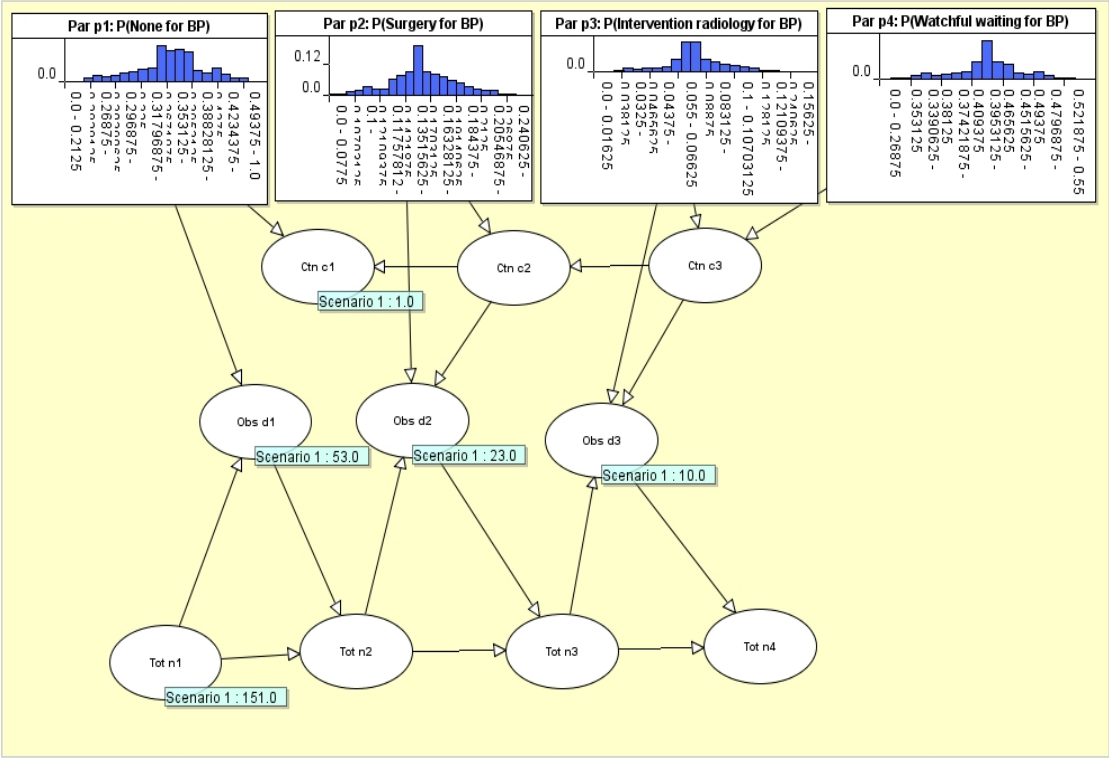


Figure 7-3 (a) A multinomial model for learning the probability of each treatment for benign pancreas

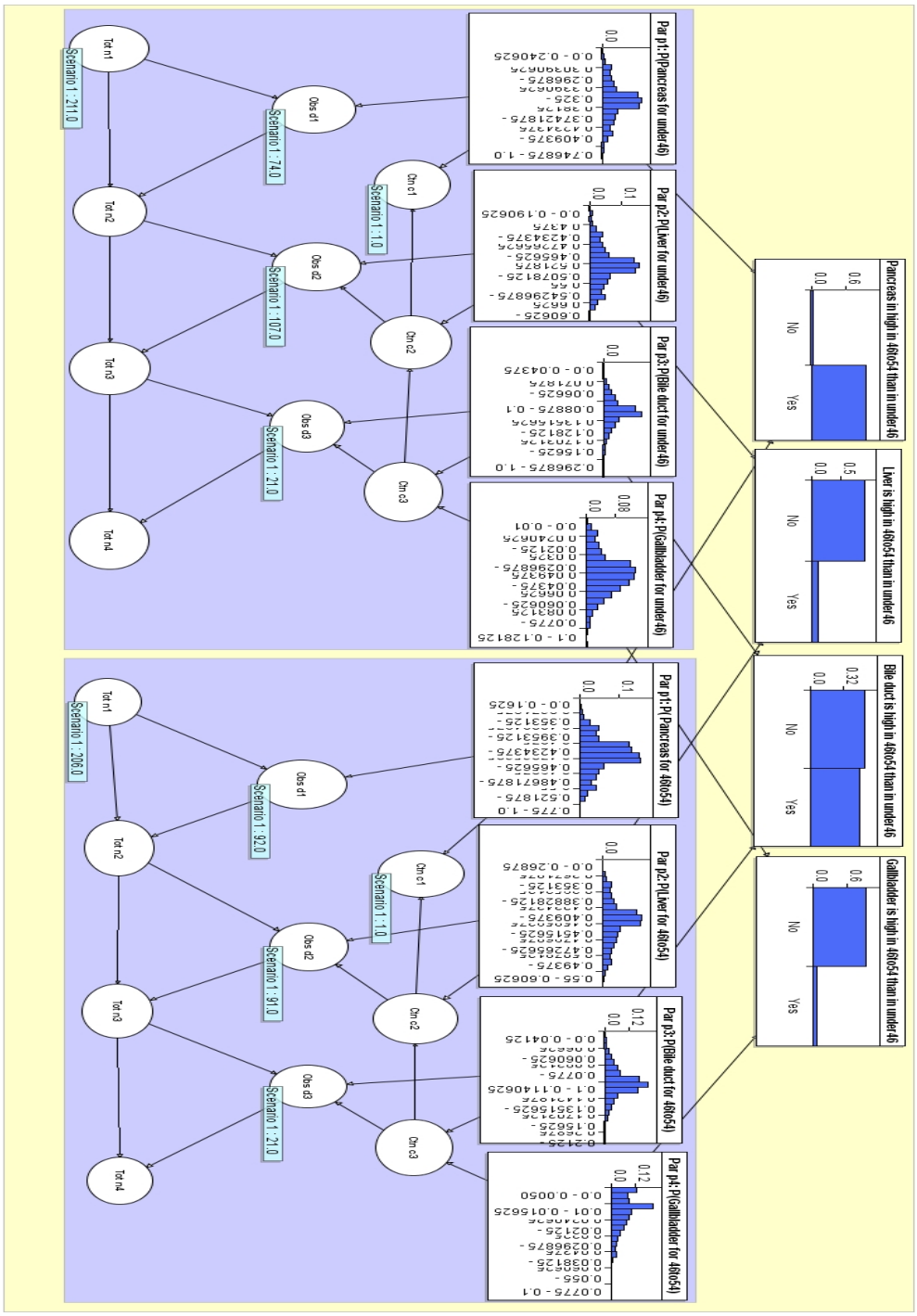


Figure 7-3 (b) A multinomial model for learning the probability of each treatment for malignant pancreas

Table 7-7 provides the posterior probability for each treatment option under each diagnosis.

Table 7-7 Probabilities over the states of Treatment given diagnoses

From	To	States	Means							
			BP	BL	BGB	MP	ML	MGB	Multiple	Unknown
Diagnosis	Treatment	Chemotherapy	-	-	-	0.08	0.13	0.06	-	0.01
		Combination	-	-	-	0.22	0.19	0.23	-	0.03
		None	0.35	0.59	0.29	0.12	0.14	0.08	0.20	0.64
		Palliative	-	-	-	0.34	0.34	0.37	0.10	0.06
		Surgery	0.15	0.14	0.49	0.21	0.15	0.24	0.50	0.07
		Intervention radiology.	0.07	0.09	-	0.01	0.06	0.02	-	0.02
		Watchful waiting	0.43	0.18	0.22	0.02	0.01		0.20	0.19

7.4 Hypothesis tests for the strength of association

The final step is to use the posterior probability distributions on the parameters of the MDT BN model, to perform hypothesis tests that provide information about the strength of each association. A hypothesis test considers two competing hypotheses and provides evidence to establish which is most plausible given the data.

Figure 7-4 shows an example of one of the networks used to test the hypotheses of interest. It can be seen in the figure that this model combines two multinomial models, each of which corresponds to the parameter estimation part for a conditional probability of one of the variables. Given the probability distribution of this parameter inferred from data, we can assess the evidence for any hypothesized difference between the true value of the conditional probability for two different states of the variables. The test is conducted by linking the two multinomial models with a boolean variable as show in Figure 7-4.

As an example, we can look for a possible difference in the probability of a particular cancerous organ given an older age group compared with the same probability given a

younger age group. The structure of a Bayesian network for evaluating this difference, for each of the organs for two age groups is shown in Figure 7-5, with four hypothesis variables. Each hypothesis test considers the following two hypotheses:

- H_1 : the probability of the organ is high for the old age group than the younger age group;

- H_2 the probability of the organ is not high for the old age group than the younger age group;

We modelled the hypothesis node for the test using the expression in Equation 7.3.

$$if\ (Pancreas\_46to54 > Pancreas\_under46, "Yes", "No") \qquad 7.3$$

Having obtained the probability of each hypothesis for the hypothesis test, the Bayes Factor (BF) is calculated for the combination. A BF is the ratio of the probability of hypothesis,H_1, being true given the data and the probability of the competing hypothesis, H_2, being true given the same data [58]. The use of a BF also allows grading this according to the scale proposed by Jeffrey in [72] (see Section 2.3).

Table 7-8 presents the results of the hypothesis tests used to analyse a possible difference in the probabilities of the different organs between the age groups under 46 and 46-54 are compared. These results derive from the hypothesis nodes of the model depicted in Figure 7-5. The value of 'Yes' in a hypothesis node is the probability of H_1 which states that the probability of the corresponding organ is higher for the 46-54 age group than under 46 age group. The value of 'No' is the probability of the competing hypothesis H_2 which states that the probability of the organ is not higher for the 46-54 age group than under 46 age group. In the Table 7-8, the row 'BF' provides the calculated ratio of the revised probabilities per hypothesis test.

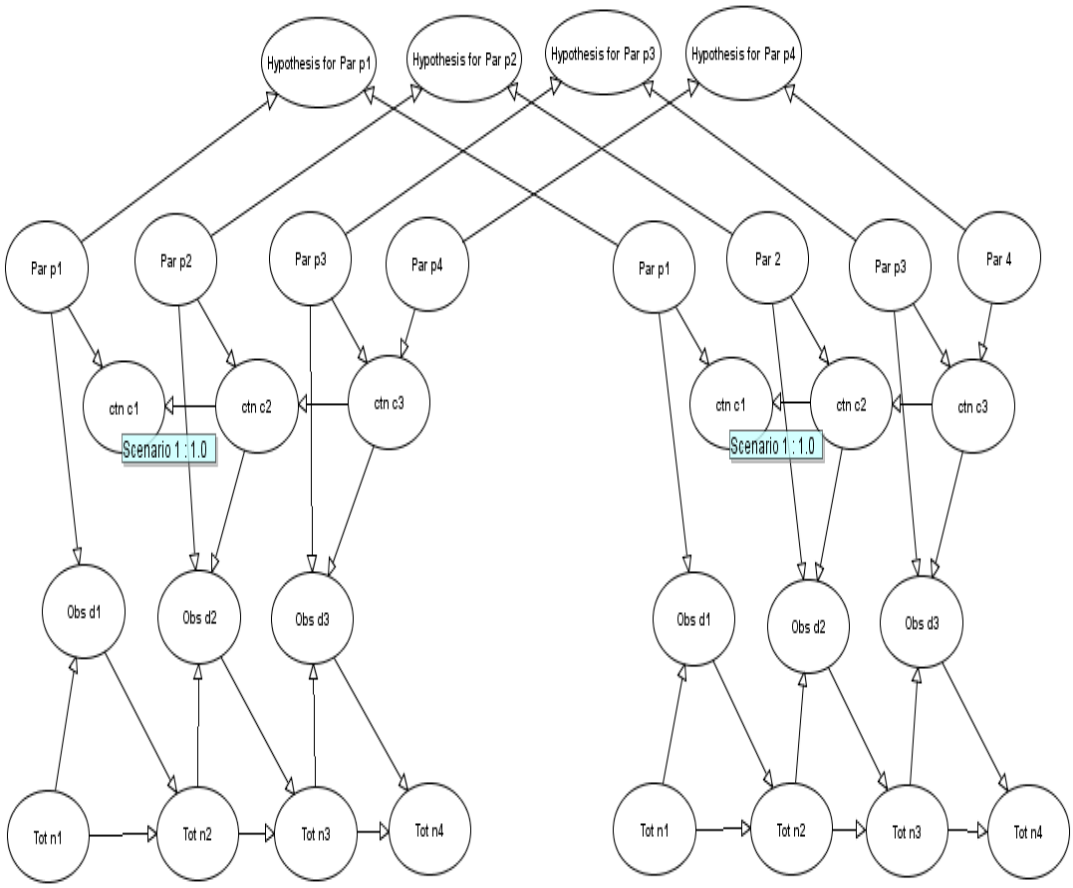


Figure 7-4 A graphical representation of the model to test the hypotheses of interest. Left-hand and right-hand sides of the graph are identical, and each corresponds to parameter estimation part for the relevant link of the BN model depicted in Figure 6-7

The results support evidence for an increase in the proportion of diseases of the pancreas for the 46-54 age group compared to the under 46 age group. In particular, the probability of H_1 is very low compared to the competing hypothesis H_2 for both the liver and the gallbladder, and the probability of H_1 is very close to H_2 for the bile duct. For the pancreas, the BF obtained is 22.45 and according to Jeffrey's scale (see Table 2-3) this is a strong evidence for H_1 that is, there is a strong evidence to suggest an increase in the proportion of diseases of the pancreas for the age groups considered

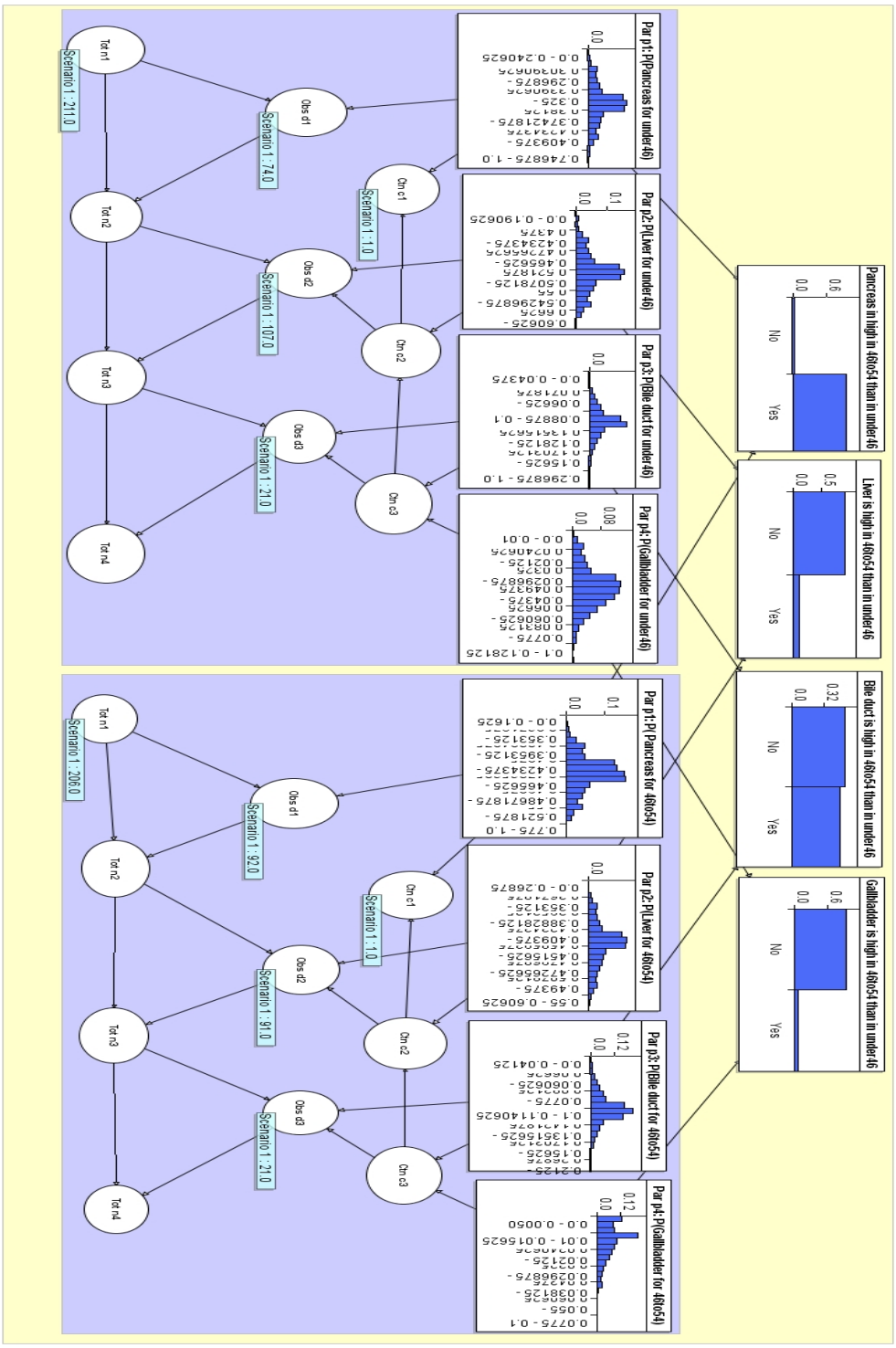


Figure 7-5 An example of actual Bayesian network to test the hypothesis of interest to the clinicians, e.g. if cancerous organs were high in an older age group (i.e., 46to54) than in a younger age group (i.e., under 46)

Table 7-8 Results per hypothesis test. A positive BF from a test indicates evidence to support an increase for the corresponding organ for the 46to54 age group compared with the under 46 age group

	Organ			
	Pancreas	Liver	Bile duct	Gallbladder
H_1	0.957	0.102	0.475	0.058
H_2	0.043	0.898	0.525	0.942
BF	22.45	0.114	0.904	0.062

Table 7-9 BF obtained for each organ from a hypothesis testing whether there is an increase for an older age group compared with a younger age group

Compared to	The probability is high for	BFs			
		Pancreas	Liver	Bile duct	Gallbladder
Under 46	46-54	22.451	0.114	0.904	0.062
	>54-60	9.115	0.034	2.143	1.759
	>60-66	1.826	0.052	7.769	2.479
	>66-71	1.598	0.038	11.06	3.557
	>71-77	0.276	0.107	23.103	4.051
	77+	1.871	0.025	15.109	4.286
46-54	>54-60	0.397	0.298	1.877	30.772
	>60-66	0.069	0.461	6.629	52.276
	>66-71	0.061	0.336	9.489	73.241
	>71-77	0.008	0.863	21.116	74.629
	77+	0.079	0.199	13.172	89.129
>54-60	>60-66	0.176	1.403	2.955	1.1
	>66-71	0.161	0.991	4.2	1.633
	>71-77	0.018	2.721	9.233	1.884
	77+	0.194	0.609	5.778	1.949
>60-66	>66-71	0.757	0.622	1.221	1.211
	>71-77	0.117	1.672	2.696	1.425
	77+	0.881	0.375	1.718	1.465
>66-71	>71-77	0.137	2.404	1.765	0.982
	77+	0.995	0.526	1.155	1.002
71-77	77+	6.104	0.19	0.552	0.848

A comparison of only two age groups is not a complete evaluation of the link *Age* → *Organ* given the available data; we therefore extended the hypothesis tests. We begin with the under 46 age group and selected each successive age group up to the >71-77

group to compare the probability of each organ for the particular age group with the probability of the organ for each older age group.

Table 7-9 presents the complete set of BFs obtained from the tests evaluating the influence of *Age* on *Organs*. The results reveal that in most cases a change from a younger age group to a corresponding older age group has (a) reduces the proportion of diseases of the liver and pancreas and (b) increases the proportion of diseases of the bile duct and gallbladder, where a BF value lower than '1' is evidence for a decrease and as above, a BF greater than '1' is evidence for an increase.

These changes are no consistent. Although reductions are more common, there are ten BFs giving evidence that cancers in the liver and pancreas have increased for older age groups. However, these BF values are considered as anecdotal evidence, revealing that no significant increase in the organ for the corresponding age groups. A similar conclusion also applies to three other BFs of the pancreas. But the BFs that resulted from the tests *'>54-60' > 'under 46'* and *'77+' > '>71-77'* are considered as substantial evidence and the BF for the test *'46-54' > 'under 46'* is considered as strong evidence to support the hypothesis.

7.5 Impact of age on other factors

In this section, and the following two sections, we give a more detailed analysis of the strength of some of the relation the MDT BN model. The main challenge of applying the technique described above is to present the information in a manner that can be understood by a domain expert.

In the MDT BN model, the variable *Age* connects *Organ* and *Type* with two links; we start, in this section, by investigating the impact of Age. In relation to *Age →Organ,* the question that clinicians in HPB Centre are interested in is whether changes have occurred in the cancerous organs according to the age of cancer patients (see Table 5.8), and in relation to *Age →Type,* the question of their interest is whether changes have

occurred in the types (Benign, Malignant and Unknown) according the age of cancer patients.

Figure 7-6 shows a simple plot of the number of patients per cancerous organ for each of the seven age groups. Using the classical statistical approach, the clinicians mainly focus on these numbers to answer to a clinical question. However, the numbers of patients for two different age groups in any of the organs are not sufficient to estimate a difference in the particular organ between the two age groups; for a reliable estimate, both the values and the uncertainty must be considered, as our method does.

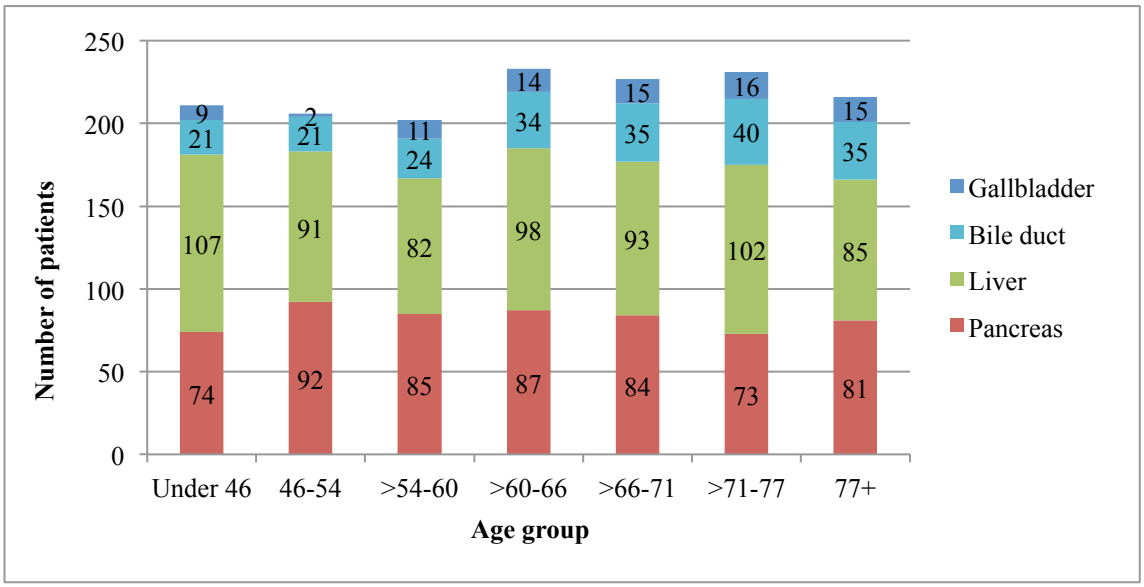


Figure 7-6 Number of patients per cancerous organ for each of the seven age groups

After calculating BFs (see Table 7-9) for the link *Age →Organ,* we analyse these values to assess: (1) changes that occur in each organ compared with other age groups in relation to a particular age group and (2) change that occurs in each organ for an older age group compared with younger age groups. Figure 7-7 (a) and (b) represent comparisons (1) and (2), respectively. In Figure 7-7 (a), we take the geometric mean of the BFs that result from the hypothesis tests, which each corresponds to one of the followings:

Evaluating Strong associations

- The probability estimated for the corresponding organ is higher for each successive older age group compared with the particular age group;

- The probability estimated for the corresponding organ is higher for the age group than each younger age group;

To calculate the BF of interest, each hypothesis test for the youngest 'under 46' age group checks (a), whereas each hypothesis test for any other age group checks either (a) or (b). For example, if our interest is in combining the BFs for the 46-54 age group in the pancreas then the hypothesis tests we carry out are: *'46-54' > 'under 46' , '>54-60' > '46-54' , '>60-66' > '46-54' , '>66-71' > '46-54' , '>71-77' > '46-54'* and *'77+' > '46-54'*.

Figure 7-7 (b) includes the followings features:

- A BF value of 1 for the 'under 46' age group. This is because the youngest age group is the starting point for comparison situation (2) change that occurs in each organ for an older age group compared with younger age groups.

- The calculated BF of the test (i.e. *'46-54' > 'under 46'*) for 46-54 age group.

The geometric mean of the BFs from 2 or more hypothesis tests for the remaining age groups. Each of these tests checks if the probability estimated for the corresponding organ is higher for the age group than for each younger age group. Thus, for the >54-60 age group we calculate the geometric mean of the BFs from the hypothesis tests: *'>54-60' > 'under 46'*, *'>54-60' > '46-54'* and *'46-54' > 'under 46'*.

Both of the graphs (a) and (b) in Figure 7-7 represent the geometric mean of BFs in a log scale. A value greater than '1' indicates evidence about an increase whereas a value which is lower than '1' indicates evidence about a decrease. In particular, Figure 7-7 (a) demonstrates that gallbladder and bile duct cancers for older age groups are higher than younger age groups. In particular, considering the calculated means, we get the highest evidence (mean of BFs is 19.08) for concluding that gallbladder cancer for the older age groups is a significantly higher proportion than for the 46to54 age group. Also we get

an almost equal strength of evidence regarding a change when bile duct cancer for each age group is compared with others. Figure 7-7 (b) further demonstrates that the bile duct cancer is consistently higher for the older age groups than younger age groups.

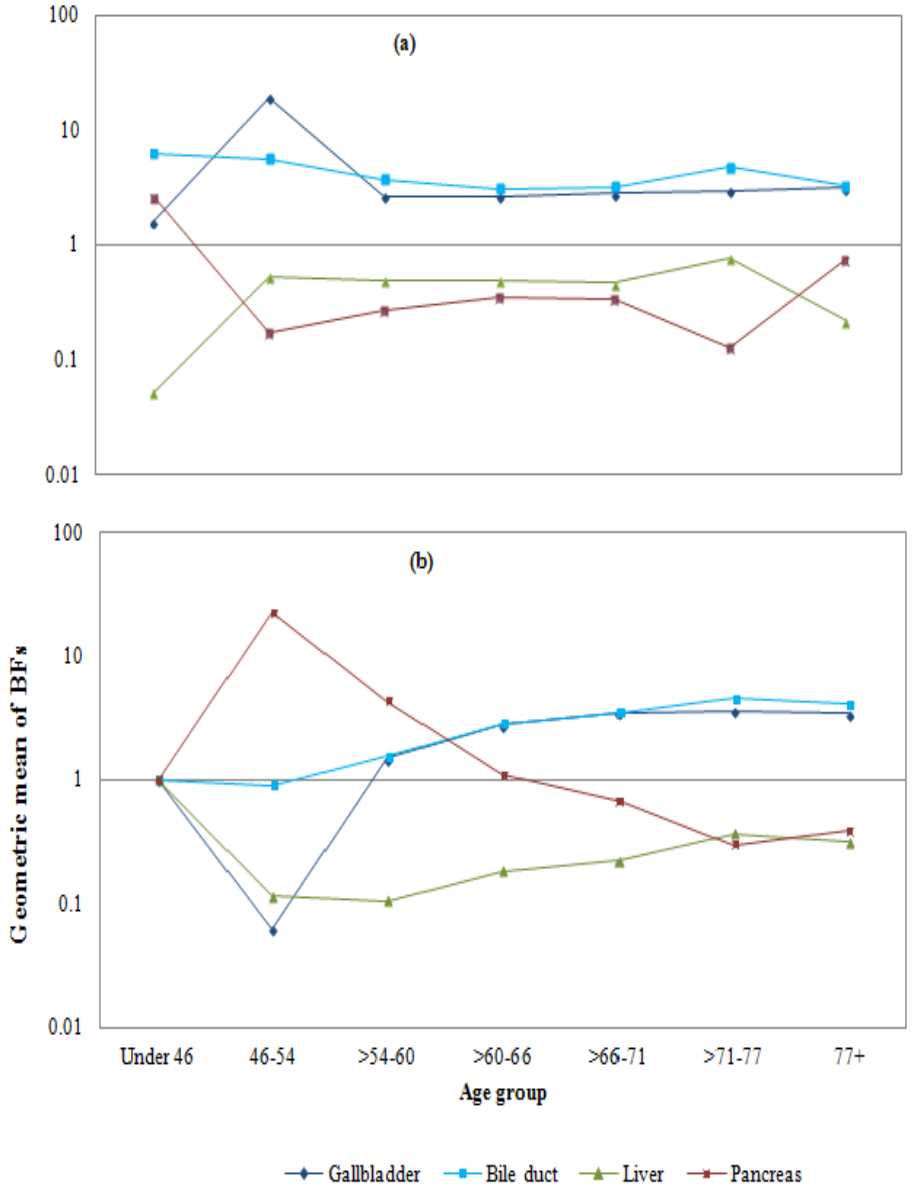


Figure 7-7 Geometric mean of BFs to assess a) changes that occur in each organ for other age groups in relation to a particular age group and b) a change that occurs in each organ for an older age group when compared with younger age groups

In comparison, Figure 7-6 makes it difficult to assess the strength of influence *Age* has on *Organs,* and consequently, to find if changes have occurred in the cancerous organs according to the age of cancer patients. But our analyses show that the link captures the impact between the variables properly. The liver and pancreas cancers decrease for older age groups (i.e. negatively influenced by the older age groups) whereas the bile duct and gallbladder cancers increase for older age groups (i.e. positively influenced by the older age groups).

Similar to the *Age* → *Organ,* we follow the same procedure to assess two comparisons for *Age* → *Type* link: (1) changes that occur in each type of cancer for other age groups in relation to a particular age group and (2) a change that occurs in each type of cancer for an older age group when compared with all younger age groups. Figure 7-9 (a) and (b) represent comparisons (1) and (2), respectively. The BFs used for the analyses are presented in Appendix B.1.

Figure 7-9 (a) demonstrates that the proportion of malignant cancer is higher in older age groups than younger age groups. In particular, we obtain the highest evidence (a geometric mean of 736.4) to conclude that the proportion of malignant cancer for the older age groups is significantly higher than under 46 age group. Also, the malignant cancer for the older age groups is higher than for the 46to54 age group (a geometric mean of 313.4). Evidence in relation to the proportion of benign cancer shows that this is lower for the older age groups than younger age groups.

Figure 7-9 (b) demonstrates that apart from the age groups 46-54 and 77+ malignant cancer is significantly higher (a geometric mean of >100) for each of the five other age groups compared to the corresponding younger age groups. In both graphs, we obtain only weak evidence for a change in an unknown cancer for older age groups compared to younger age groups.

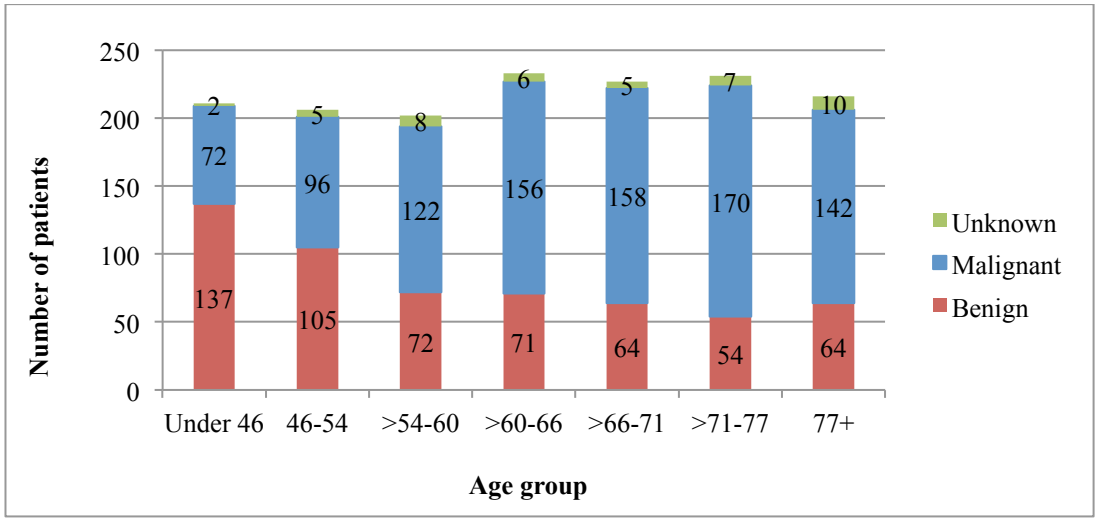


Figure 7-8 Number of patients per cancer type for each of the seven age groups

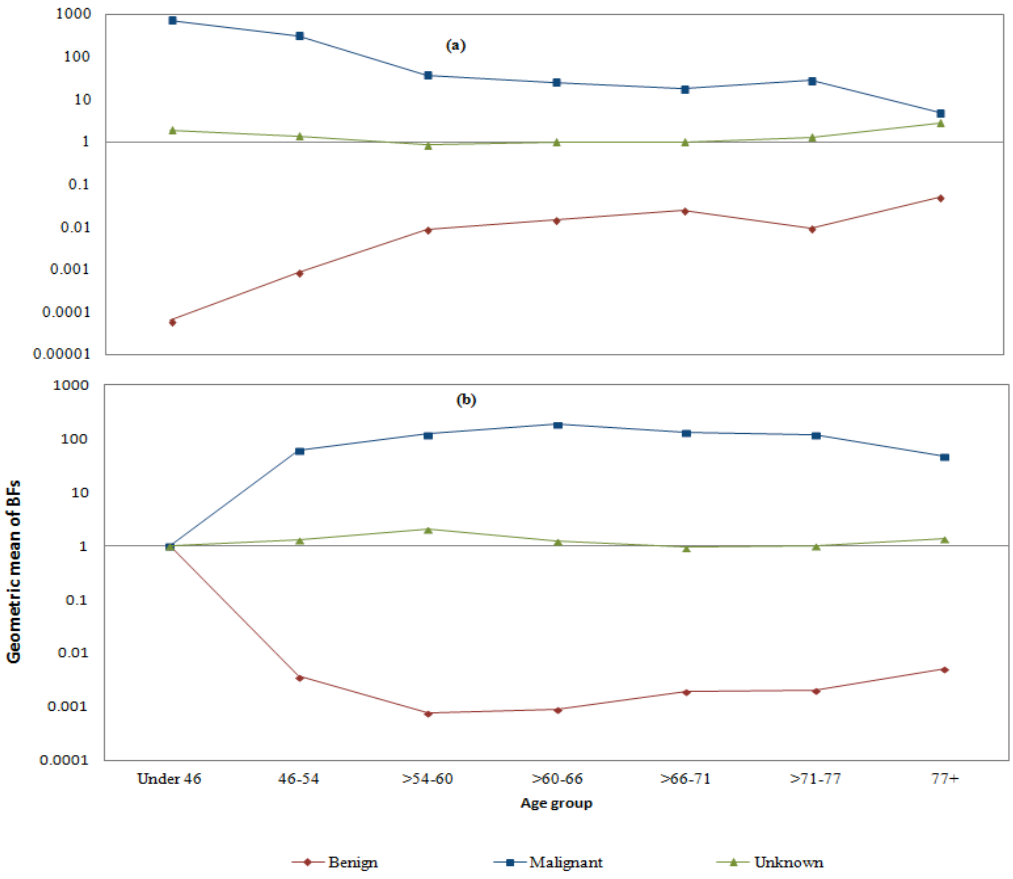


Figure 7-9 Geometric mean of BFs to assess a) changes that occur in each type of cancer for other age groups in relation to a particular age group and b) a change that occurs in each type of cancer for an older age group when compared with younger age groups

7.6 Efficiency in the MDT meetings with years

This section checks if the MDT meetings are becoming efficient with years by assessing the link *Year → Number of meetings*. A number of hypothesis tests are carried out to find evidence about the changes in each number of meeting category, from one year to another.

Table 7-10 presents the Bayes Factors from the hypothesis tests for meeting categories 1 to 4 and more, for the five years considered. The BFs mostly indicate that single MDT meetings have decreased and multiple meetings (i.e. 2, 3, 4 or meetings) have increased over the years. However, even these results do not clearly provide an answer to the question "Given the data, how strong is the evidence that the proportion of multiple MDT meetings was greater for 2008 than all of the three earlier years?" To understand the change better, we do a further analysis on the results. The main point of interest is to know what change occurs in each meeting category for one year compared with the corresponding earlier years.

Table 7-10 BFs (with two decimal places) that derive from the hypothesis tests for meeting categories 1 to 4 and more, for the five years considered

Compared to	The probability is high for	BFs			
		1	2	3	4 or more
2005	2006	18.78	0.03	2.59	4.41
	2007	0.67	0.36	3.74	6.48
	2008	0.03	7.58	3.69	1.68
	2009	0.04	3.85	2.65	18.89
2006	2007	0.03	22.43	1.04	1.07
	2008	0.01	51.07	0.97	0.31
	2009	0.01	48.65	1.02	3.6
2007	2008	0.04	22.7	0.78	0.22
	2009	0.05	8.19	0.88	3.47
2008	2009	0.44	0.8	0.96	11.62

As before, the graph shows the following:

- A BF value of 1 for 2005, as this year is the starting point for investigating changes in the proportion of multiple meetings.

- The BF from the hypothesis test *06>05* for 2006.

- The geometric mean of the BFs results from the hypothesis tests for 2007 to 2009. Each of the hypothesis tests checks if the probability estimated for the particular meeting category is (a) higher for the year than for each previous year or (b) higher for each previous year (when investigating about 2008, the years we look at are 2006 and 2007) than for of that year's predecessors. Thus, for 2008 we calculate the geometric mean of the BFs that accumulate from the hypothesis tests: *08>05*, *08>06*, *08>07*, *07>06*, *07>05* and *06>05*.

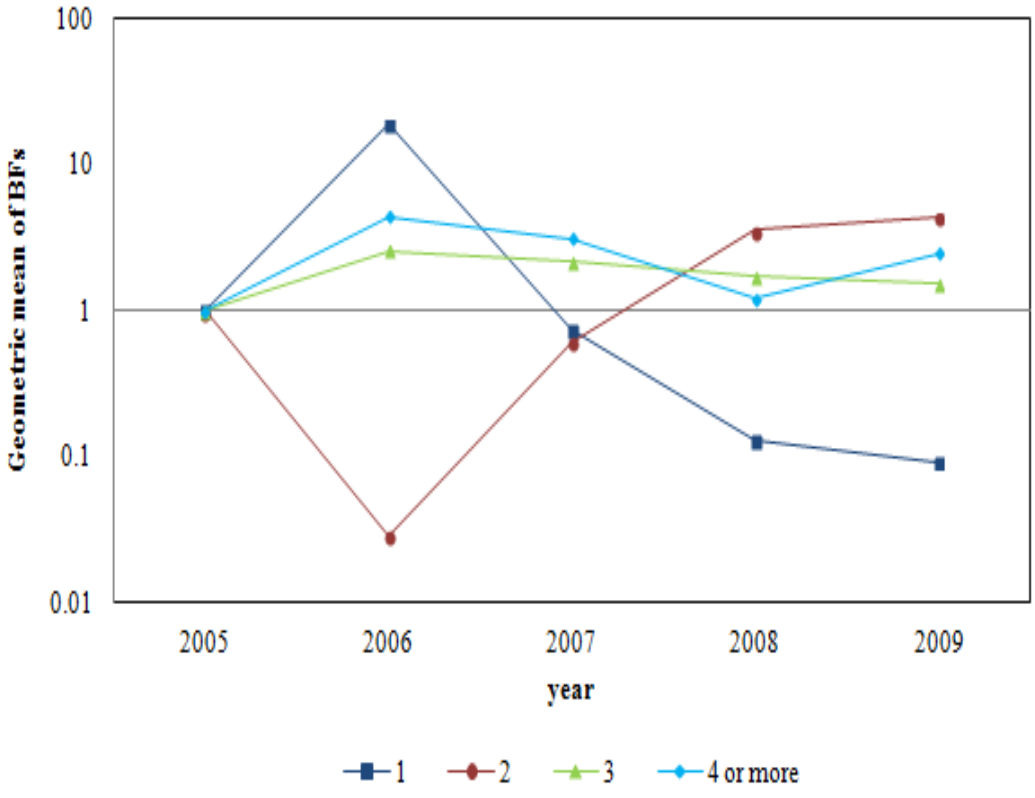


Figure 7-10 Geometric mean of BFs to assess efficiency in the MDT meetings with years

Figure 7-10 presents the meeting categories by four different lines. It illustrates that multiple meetings (twice in particular) increase for the later years (i.e. 2008 and 2009) compared with the previous years. For 2009, the evidence (mean of BFs = 4.36) was slightly stronger for more multiple meeting than for 2008 (mean of BFs = 3.55). The figure also illustrates that the proportion of other multiple meeting categories (thrice and four or more, in particular) increased for each later year when compared with the corresponding previous years. But the proportion of single meetings was only higher for 2006 than 2005, otherwise decreasing a) for 2007 than 2006 and 2005, b) for 2008 than 2007, 2008 and 2009 and c) for 2009 than all the previous years.

Given a belief that the organisation of MDT meetings had become more efficient over the study years, the clinician (our expert) has assumed the data would show evidence of fewer multiple meetings for later years. However, the analysis shows that the changes in the proportion of the number of meeting categories were the opposite of that assumed by the clinician; however, the evidence for an increase in the proportion of multiple meetings is not strong.

7.7 Impact of year on type, organ and treatment

This section evaluates the other links from *Year* in the MDT BN model; apart from *Number of meetings,* the variable *Year* connects to *Type*, *Organ* and *Treatment* in the MDT BN model. The results derived from the hypothesis tests carried out to assess these links can be found in Appendix B.2. Figure 7-11 (a), (b) and (c) demonstrate changes to the values of the variables, as captured from the estimated probabilities, for a later year compared to the relevant previous years. Figure 7-11 (a) reveals that the proportion of malignant cancers was higher for 2007 than for 2006 and 2005 (mean of BFs = 5.68) but less for both 2008 and 2009 than for the previous years. In contrast, for both 2008 and 2009 the link has captured an increase in the proportion of benign cancer against the corresponding previous years.

It is an interesting fact that the findings regarding benign disease confirm the clinician's expectation. Since the variable 'year' in the MDT BN represents changes from any unknown influential factors over time, the clinician view's that progress has been made in detecting more cancers at a benign stage would be shown as an increase in the proportion of benign disease in later years. However, we note that the evidence suggesting an increasing proportion of benign disease in the later years (2008 and 2009) is very weak (geometric mean of relevant BFs is ≥ 3).

Figure 7-11 (b) reveals that up until 2007 the proportion of liver cancer for later year was more than the corresponding proportion for previous years. The proportion of liver cancer reduced for 2008 (mean of BFs = 0.26) and became high again for 2009 (mean of BFs = 1.23). As in the liver cancer, the derived evidence indicates that the proportion of gallbladder cancer for 2009 was higher than the previous years (mean of BFs = 1.92). In contrast, the evidence about bile duct and pancreas indicates a reduced proportion of these cancers for the later year. However, in each case there is only a week evidence to support these changes.

In relation to the variable *Treatment*, analysis is performed to check the impact of *Year* on four outcomes: combination, chemotherapy, palliative and surgery. As illustrated in Figure 7-11 (c) for 2008 recommendations regarding palliative care and combination reduce compared to the previous three years (means of BFs are 0.42 and 0.02, respectively). The analysis also reveals that the MDTs are recommending more surgery to patients. Means of BFs that derive for 2007 and 2008 were 7.84 and 6.16, respectively. These findings examine the details of the link capturing association between *Year* and *Treatment*. Improvements within MDTs in higher years may have made it easier to recommend surgery as a treatment option than before.

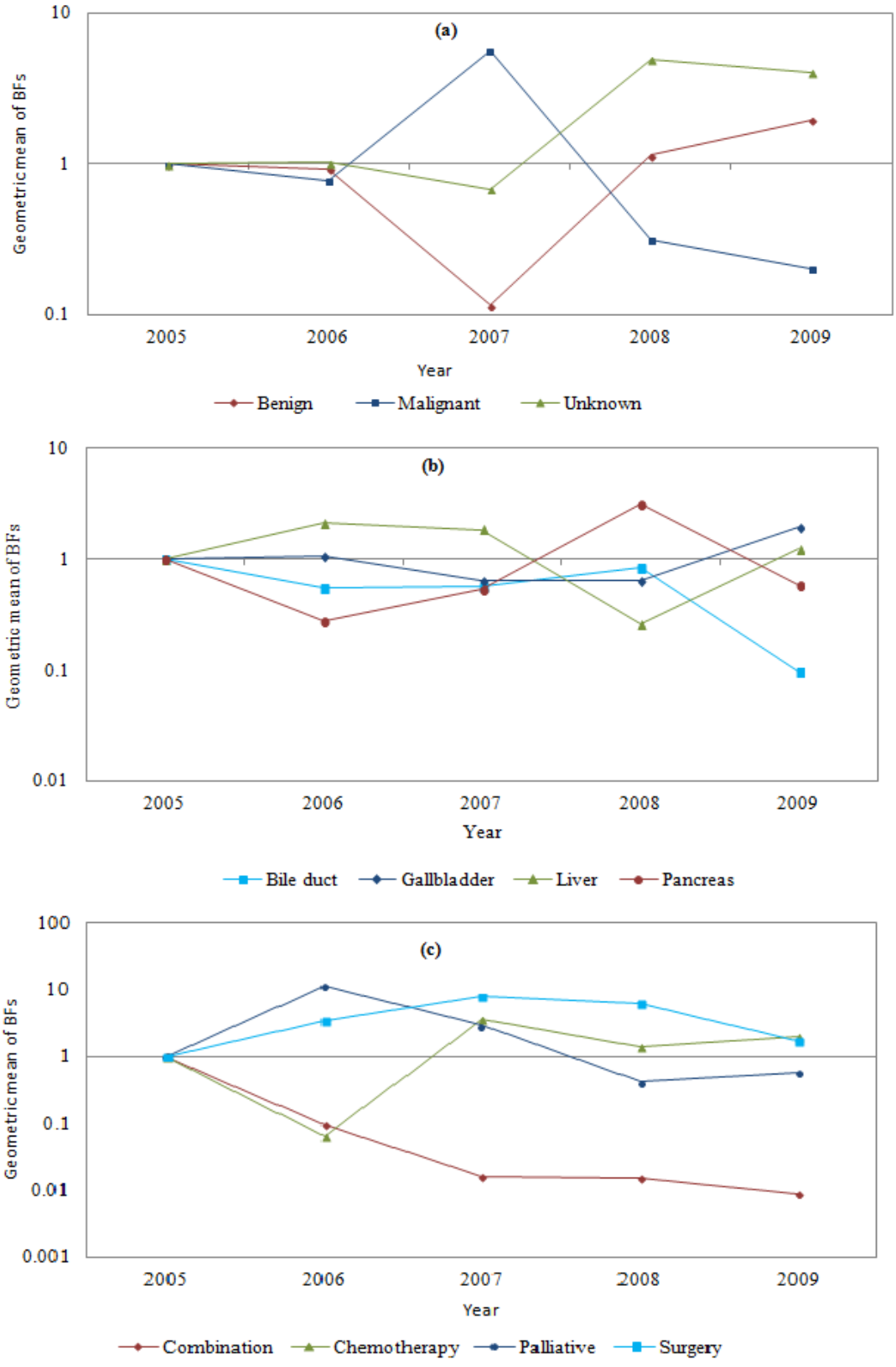


Figure 7-11 Changes with years a) in Type, b) in Organ and c) in Treatment

7.8 Impact of diagnosis on the treatment and number of meetings

This section evaluates the impact of *Diagnosis* on the *Treatment* and on the *Number of meetings*. The analyses performed basis the hypotheses on the expectations and clinically interests of the clinician. The hypothesis suggested by the clinician to explain the link *Diagnosis* → *Treatment* is that the treatment surgery has been recommended more in situation when the diagnosis of cancer was complex (see Table 5-10). The hypothesis suggested for explaining the link *Diagnosis* → *Number of meetings* is that more multiple meetings were held in situation when the diagnosis of cancer was complex (Table 5-10). The aim of the analyses is therefore to find out whether the relations correspond to the suggestions made by the clinician.

Similar to the other sections, the analysis is based on hypothesis tests. With each of these tests we check:

- In relation to the probability estimated for an outcome for one diagnosis, the probability of the same outcome is higher for each successive diagnostic category;

- The probability estimated for an outcome for one diagnosis is higher than the probability of the same outcome for each of the diagnostic categories that appear earlier;

For benign pancreas, we mainly carry out a test to check (a), whereas for others a test checks either (a) or (b). The BFs resulted from these tests are provided in Appendix B.3. Figure 7-12 (a) and (b) present the geometric means of these BFs to show the strength of evidence for the differences that occurs in each outcome of the variables for each diagnosis category compared to others.

In relation to the variable *Treatment*, we obtain the strongest evidence to show that the recommendation of surgery is less for unknown diagnosis than other known diagnosis

categories. The mean BFs of 329.95 and 36.6 also suggest that the surgery recommendations for other diagnosis categories were significantly higher than benign liver and benign pancreas, respectively. In relation to the variable *Number of meetings*, we obtain evidence to show that multiple meetings (twice and thrice in particular) for other known diagnosis categories are more likely than for benign liver (means 3.75 and 1.5 for twice and thrice, respectively). Figure 7-12 (b) also reveals that single meetings are more likely for malignant diagnoses than benign liver.

As before, these findings help to explore the characteristics of the impact of *Diagnosis* on the *Treatment* and *Number of meetings* variables captured in the MDT BN.

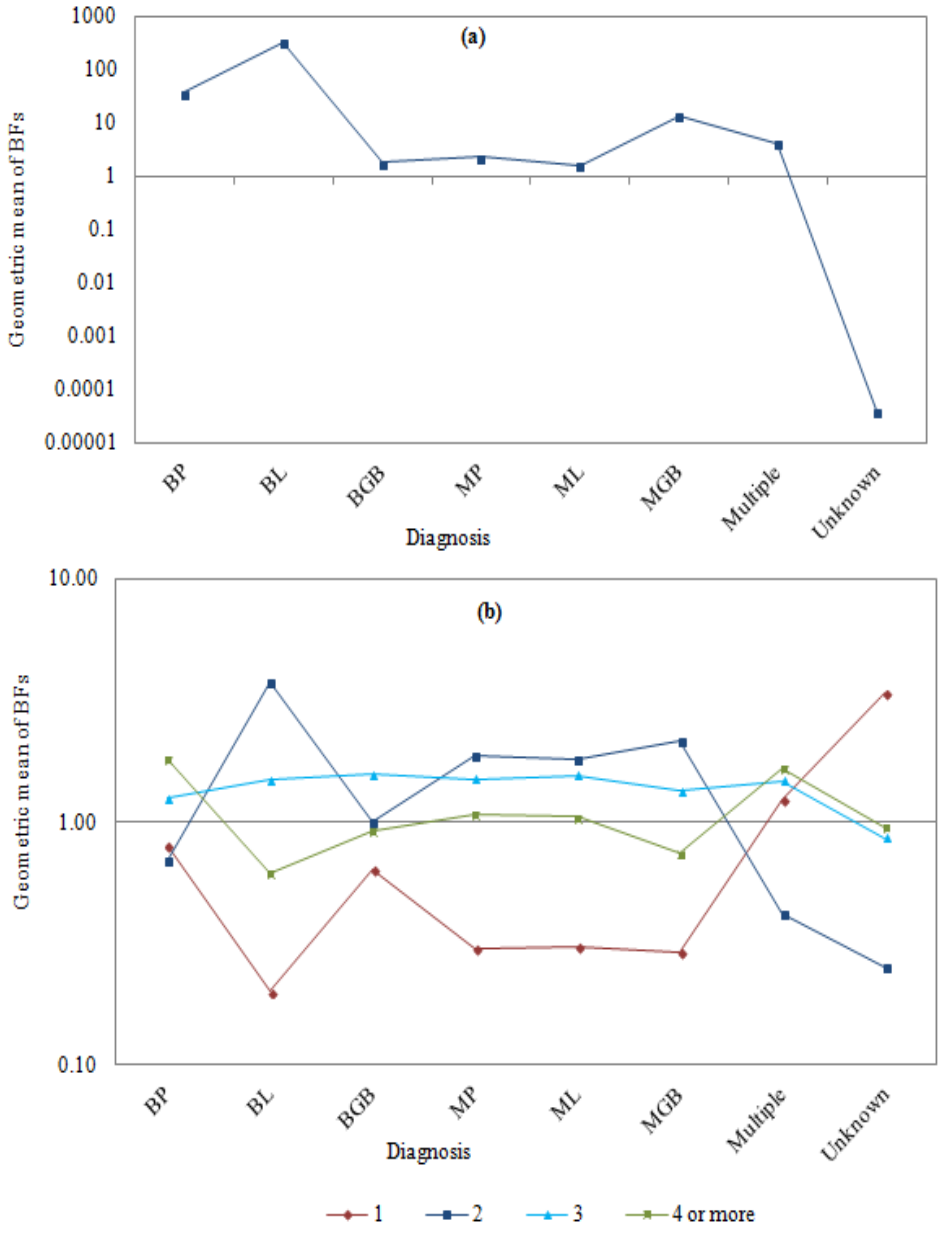


Figure 7-12 Evidence for a difference that occurs a) in surgery and b) in each meeting category for each diagnosis category compared to others

7.9 Using MDT findings for health care management

Based on the findings from the above analyses it is possible to generate the following recommendations for the management:

- The study showed how the age of a patient could help to understand which organ is likely to carry the infection and what is the likely stage of the disease. With the help of this information health professionals can refrain from arranging unnecessary tests while working towards diagnosis.
- An earlier indication of the disease help health professionals to know about the treatment required, and thus allowing the management plan ahead the resources. For example, for a patient with a possibility of a benign cancer the possibility of receiving chemotherapy as a treatment recommendation is less but receiving a palliative care is more.
- The study showed no evidence of an increase in more than 2 meetings per case with times. This suggests that while holding meetings twice to discuss a case is likely to show benefits in providing better treatment recommendations, the use of more meetings in most cases is not necessary.

7.10. Summary

We have presented a novel approach to evaluate the strength of a link in a BN model, using data. The strength is measured using Bayes Factors comparing hypotheses about the true values of conditional probabilities of the variables connected in the model. By analysing the results of such tests in the MDT BN model, we are able to answer queries about the MDT process. For example, although the expert has assumed that the data will show a decrease in single meetings with years due to improvements with the MDT process, we do not find strong evidence to provide support for this.

Chapter 8

A Method for Modelling Associations

Chapters 5, 6 and 7 demonstrated how data and knowledge can be used together to construct a probabilistic model representing strong statistical associations between the domain factors. The method used is solely based on a health service case study. In this chapter we discuss how this can be applied in other domain and what would be required to test the generalizability of the method. The chapter is organised as follows: Section 8.1 describes the modelling steps, along with the assumptions made and issues found. Section 8.2 provides details of a comprehensive tool based on the study method and the possible test for checking generalizability, Section 8.3 discusses some related research studies and Section 8.4 presents the summary.

8.1 Modelling Associations

This thesis considers both knowledge and data for modelling strong associations. In particular, three issues have been tackled. First, how knowledge and data can be used to identify if there exist associations between variables, secondly, how a complete structure of statistical associations can be build, and finally, how the strength of each association modelled within the structure can be assessed to address queries regarding the problem domain. Figure 8-1 depicts the process and the brief description about each step is as follows:

- Variables are selected from data. These variables represent the main factors for modelling a BN.

- The knowledge of experts is used to determine four relation types, namely 1) illegal, 2) definitional, 3) impossible and 4) hypothetical, between the variables considered. This is done by presenting the variables in a tabulate form, such as Table 5-7, which allows experts to identify the relation types that may exist from one variable to others.

- The structure of a BN is modelled based on the relations identified in step 2.

- The fragments of the BN, each of which represents causal relationships assumed by the expert between a variable and its parents, are checked against data. For each fragment, we present hypotheses for the different combinations of parents. If the hypothesis corresponding to all the suggested parents receives the highest evidence of support from the data, the fragment remains the same; otherwise it is revised according to the data supported hypothesis

- Based on all revised fragments found from step 4, the structure of the initial BN is revised. The definitional fragments, in which the relations between the relevant variables are definitional, cannot be checked against data, so we do not make any change within these fragments.

- Multinomial models are created to learn the CPT for every relation that the expert considered as a causal relation and later received the support of existence from the data. The CPT contains the probabilities of the states of a variable under its parents' states. We continue to learn probabilities for each state of the parent separately. Firstly, we count the frequency of occurrences of the states under the particular state from data, and then we create a multinomial model to calculate the probability distribution over the conditional probabilities by Bayesian inference. The process continues until all the states of the parent have been considered.

- The strength of each association is then assessed using data. Once the distributions over the conditional probabilities are calculated, these values needed to be used to evaluate the strength of associations between variables. We

therefore examine differences in probabilities between different states of the parent by means of a set of hypothesis tests, by combining two multinomial models.

- Further analyses on the data that are found the assessment are performed. In order to understand any existence of causality if there are changes in the ordered child variable due to changes in the parent variable. When the states of the child are not ordered, then some meaningful hypotheses are used to describe how the variable may influence on its child to investigate a change.

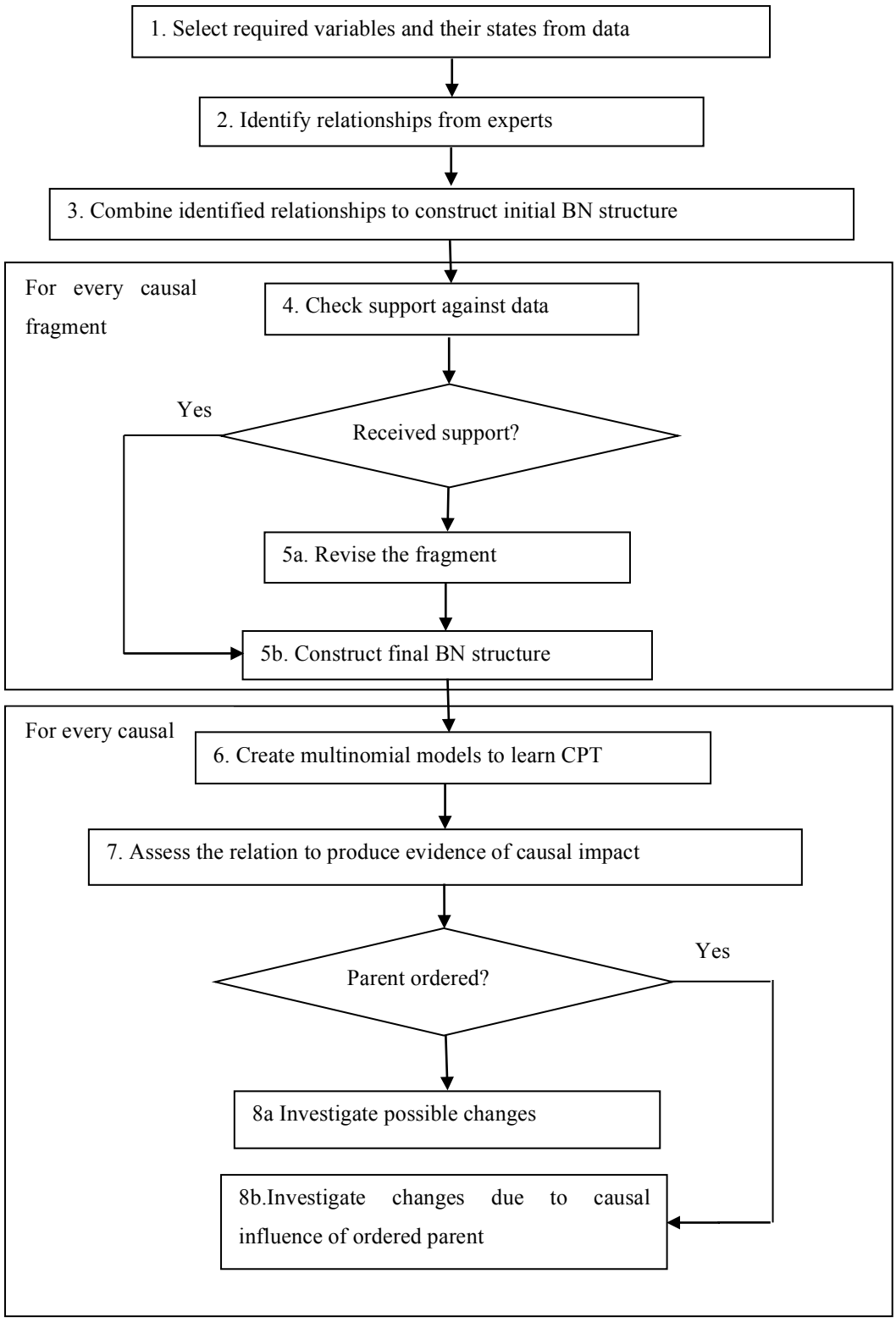


Figure 8-1 A flow chart of modelling associations

8.1.1 Assumptions

The steps mentioned earlier are used for modelling statistical associations under the following assumptions and limitations:

- Variables are identified in two ways: (a) by analysing the case study data and (b) by using the values of other variables. Therefore, it is not required to find variables from any related literature.

- The data used don't have any missing value. We removed inconsistencies that existed at the beginning. This implies that the techniques used do not consider any imputation of missing values.

- The method of modelling association and making sense of causal influence depend on the use of continuous nodes. The AgenaRisk tool has the ability to properly incorporate continuous nodes and does not require using static discretisation [177]. It uses the dynamic discretisation method proposed in [79] which produces results with greater accuracy

- The BN constructed by taking account of the order in which the variables appeared in the problem domain. The term causality only used for the relations considered in the expert BN model. However, the direction of a relationship defined within the expert BN remains unchanged at every stage of the modelling process.

- In learning the structure of a causal fragment, $Uniform\,(0,1)$ distribution is used as the prior belief on each parameter. This was to demonstrate that before the data is observed all the points of the parameters remains equally likely. However, one could also set a non-uniform distribution based on some prior values and this will not provide any restriction to learn the existence of a fragment by following the method used in this thesis.

- Similar to the structure, the prior belief on the parameters of each multinomial

model was modelled as $Uniform\ (0, 1)$. A multinomial model could also be build by considering a non-uniform prior distribution for its parameters if experts have some values in mind. This selection will not restrict calculating probabilities after finding relevant observations. If the number of trials is observed from data, one will yet be able to calculate the posterior distribution of the parameters by means of Bayesian inference. The work in [178] demonstrated this.

8.1.2 Modelling issues

Input data source

The domain expert's knowledge and historical data containing details about the important factors and their values are the input data sources.

Causal fragment

The expert BN in Chapter 5 only modelled causal relations between variables. All causal assumptions regarding relation s and their directions are made based on the knowledge of the expert. The use of the data only explained if any association between the variables is plausible or not.

The method used for learning the plausibility of relations within a causal fragment relies on manual calculation of the joint probability and normalised score. Figure 8-2 shows the time currently requires for calculating parameters from the observed data. As can be seen from the figure the time requires increases in a rapid manner with an increase in the number of observations per parameter. We can reduce this time by increasing the size of the RAM space of the machines used.

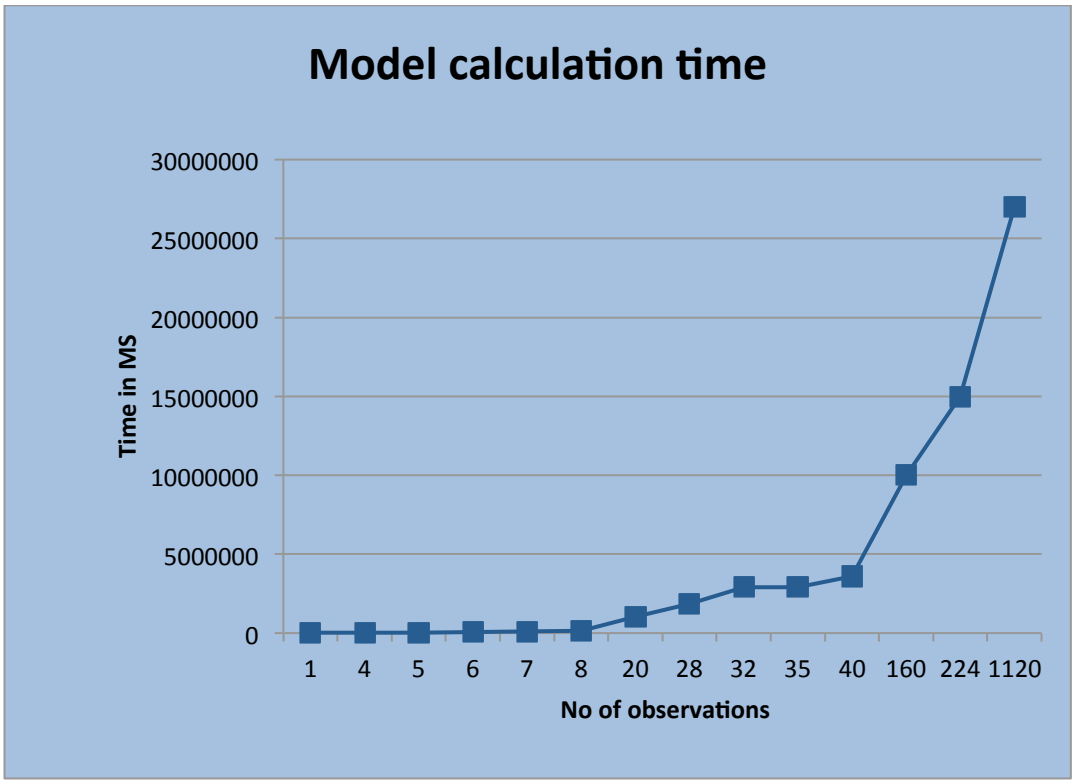


Figure 8-2 Time needed to calculate a parameter model

Learning the strength of associations

The method used for learning the strength of a plausible relation initially calculates the conditional probabilities for given states of the parent, and then the BFs by comparing the probabilities calculated. Similar to causal fragments the time requires for learning increases with the number of the states of the child variable. We analyse the BFs to bring evidence in support of a predefined hypothesis. The user must have the expertise to define a meaningful hypothesis.

8.2 Outline of a comprehensive tool

In this section, we illustrate how the method proposed in this study can be applied to assess the quality of food in the food industry by an inspector. For an input dataset a set of variables will be identified as important variables by the tool implementing the method. The structure of the causal BN based on the knowledge of the inspector is shown in Figure 8-3(a). To construct this model the inspector enters details about the causal direction between

- *Season* and *E Coli*
- *Season* and *Level of danger*
- *E Coli* and *Level of danger*
- *Level of danger* and *Food quality*
- *Packing* and *Level of danger*
- *Packing* and *E Coli*

The final model in Figure 8-3 (a) shows no association between *Packing* and *E Coli* after investigating the causal fragment *Season* $\rightarrow$ *E Coli* $\leftarrow$ *Packing,* and no association between *Packing* and *Level of danger.*

Figure 8-3 (b) illustrates how the tool assess the strength of each relation of the final BN model. At this stage, the instructor provides the counts of the states of the child variables for the states of parents as an input data. The output is a set of Bayes Factors (BFs) each of which calculates from a joint multinomial model. Constructions of these auxiliary BNs stay hidden.

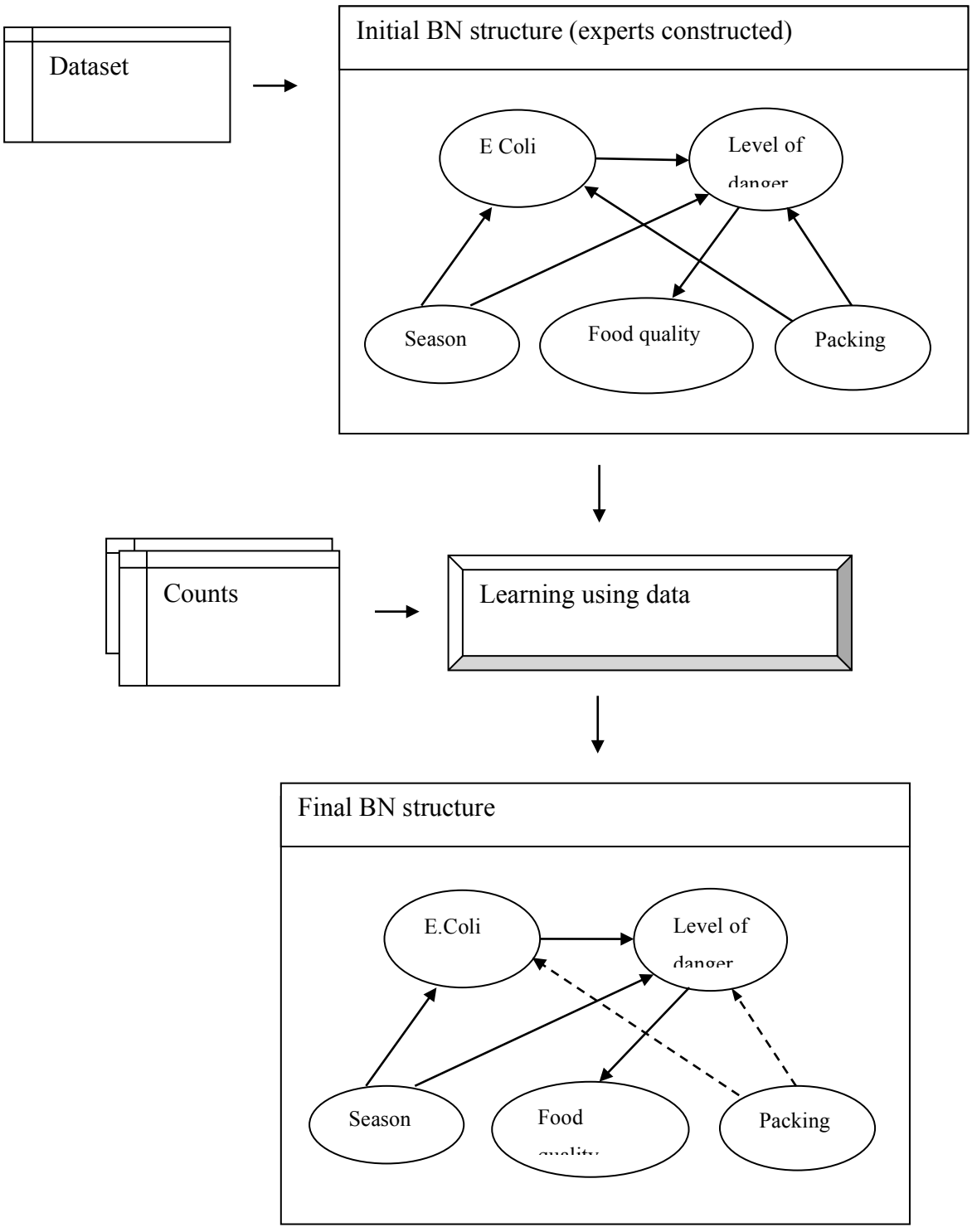


Figure 8-3 (a) How to construct a BN structure using knowledge and data

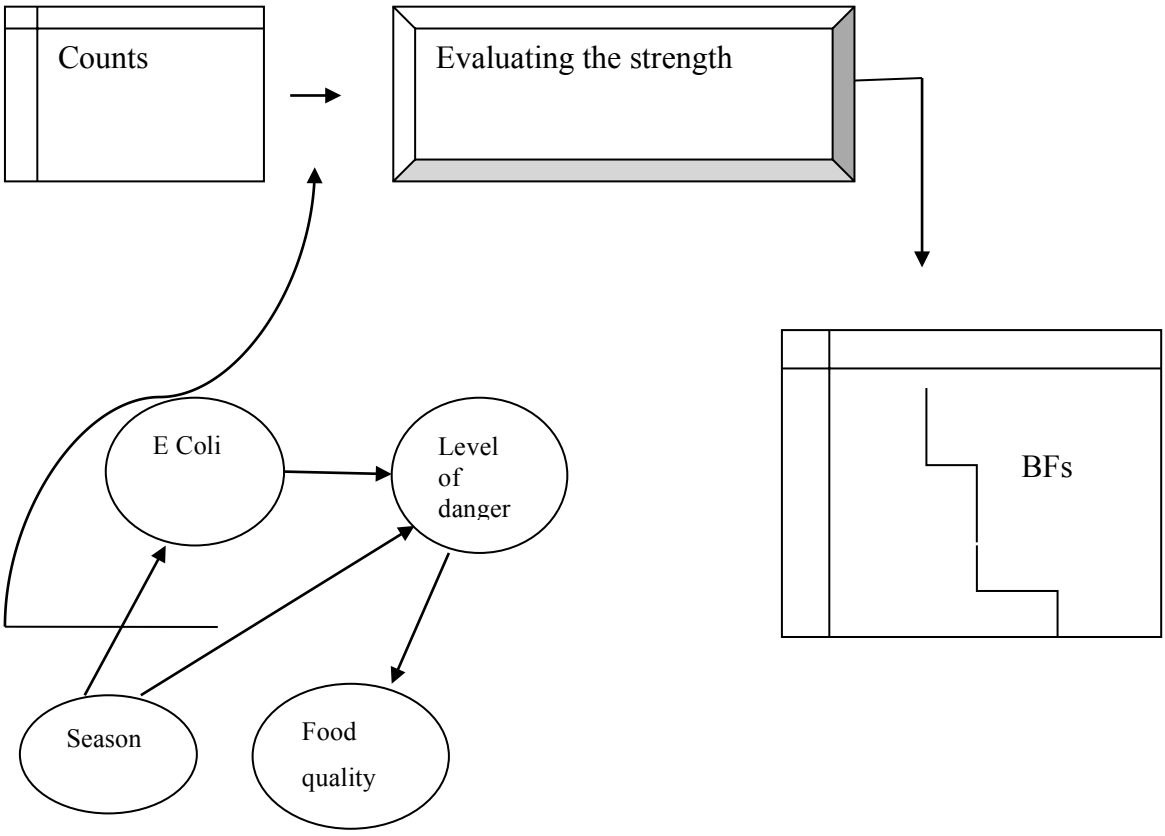


Figure 8-3 (b) How to evaluate the strength of a relation and address relevant queries

The clinician can also analyse the produced BFs to obtain meaningful information about the domain. It is believed that an ordered variable often introduces changes in ordered way, and for other discrete variables this may not be the case.

Table 8-1 BFs to assess changes in food quality according to the level of danger

Compared to	**The probability is high for**	**BFs of lipid disorder**		
		High	Medium	Low
Low	Medium	2.63	1.44	0.2
	High	7.94	4.03	1.2
Medium	High	7.17	0.73	0.8

Let us assume that, in case of *Level of danger* → *Food quality* of the example model, the variable *Level of danger* is an ordered variable and thus, the quality of decreases as the level of danger changes from low to high. Figure 8-4 is an analysis of the BFs presented in Table 8-1. The graph is produced on a log scale by taking BFs (2.63, 7.94 and 7.17) as inputs. The points on the graph are generated as follows:

- For Low → 1
- For Medium → 2.63
- For High → $the\ geometric\ mean\ of\ BFs$ (2.63,7.94,7.17)

$$= \sqrt[3]{2.63 \times 7.94 \times 7.17} = 5.31$$

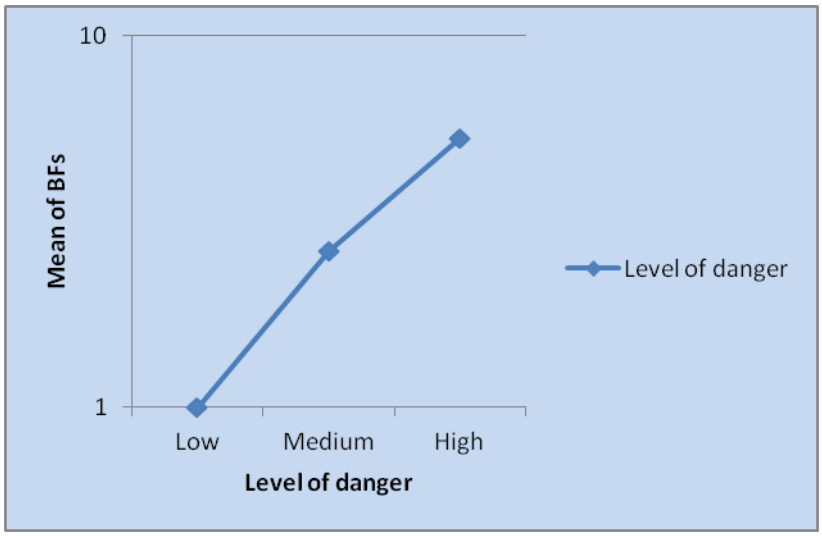


Figure 8-4 Changes that occur in the quality of food with the level of danger

If the states of the variable *Level of danger* are not in order, the variable does not necessarily indicate that the quality of food decreases gradually with the level of danger. Therefore, to analyse an impact we have to choose which assumption is best to follow. Consequently, Figure 8-5 is produced to check if for the higher probabilities of both the medium and high danger level, the probability of low food quality increases. Similar to

the earlier analysis, the graph is produce on a log scale by taking BFs (2.63, 7.94 and 7.17) as inputs. The points on the graph are generated as follows:

- $For\ \ Low \rightarrow the\ geometric\ mean\ of\ BFs\ (2.63, 7.94)$

$$= \sqrt[2]{2.63 \times 7.94} = 4.57$$

- $For\ Medium \rightarrow thegeometric\ mean\ of\ BFs\ (2{,}63, 7.17)$

$$= \sqrt[2]{2.63 \times 7.17} = 4.34$$

- $For\ High\ \rightarrow\ the\ geometric\ mean\ of\ BFs\ (7.94, 7.17)$

$$= \sqrt[2]{7.94 \times 7.17} = 7.55$$

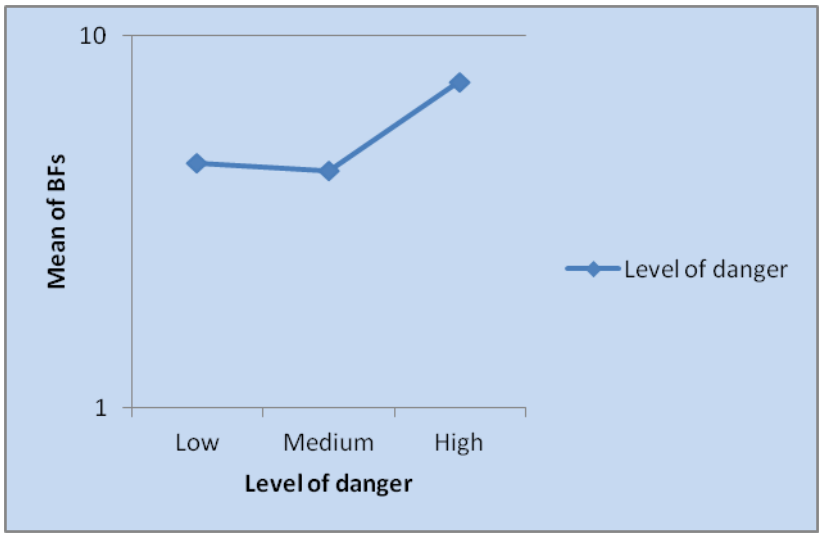


Figure 8-5 Changes that occur in the low quality of food for the Medium and High danger levels compared to the Low level

8.3 Discussion and related research

This section discusses some studies in which the authors have performed works that are similar to what we investigated for this thesis. The interest here is to show that the methodology used in this research could be applied to other studies.

8.3.1 Medical case studies with related aims

Ogunsakin et al. [179] tried to find out if age and blood cholesterol have any effect on the blood pressure. The study was based on data obtained from a teaching hospital. The type of the studied variables was categorical, and in order to test the level of significance between these variables they performed multiple regression analysis. After obtaining the regression equation the regression coefficient, R^2, is used to explain a possible variation such as if blood pressure increases as the age increases, thus indicating age is significant in the change of blood pressure.

Tomera and Harakal [180] investigated what causes high blood pressure using animal data. Similar to the above study, they performed multivariate regression analysis to determine about relationships between blood pressure, hypertrophy, calcium and cadmium. They used the square of the correlation coefficient, R^2, the adjusted R^2 values, and the power statistics to test the existence of a relation and to understand how one variable influences another.

Sultan and Rao [2] investigated the possible association between osteoporosis and periodontal disease. Their research was based on data from 80 postmenopausal women. They assessed the relationships between variables by Pearson's correlation coefficient. The correlation coefficient, whether positive or negative, determines the degree or extent of association between the two variables; a large value suggests the relationship may be significant, whereas, a small value indicates that it might have appeared by chance. Table 8-2 displays the main results of their study.

Table 8-2 The results produced in [2]

Variables	Correlation with BMD	
	r	p-value
Bone mineral density (BMD)	...	...
Age	-0.381	0.000
Menopausal Age	0.116	0.307
Years since menopause	-0.417	0.000
Body Mass Index	0.263	0.018
Plaque index	0.027	0.814
Gingival index	-0.126	0.265
Clinical attachment loss	-0.009	0.937
Alveolar bone loss	-0.093	0.410
Number of remaining teeth	0.047	0.047

In [181] Rona et al. checked if two variables are significant at the 5% level. Based on the study data, containing blood pressure information of 9 years old children, they concluded that children who had a low birth weight had high systolic blood pressure ($p < 0.05$), but not diastolic blood pressure, and those who had a shorter length of gestation had high systolic blood pressure ($p < 0.01$).

These studies have focused on evaluating relations using data derived in observational studies. They reported results to suggest about the existence of correlations between variables. However, the use P-values for assessing the strength of relations mean a decision maker will require making his own interpretation of the measures. The resulting decision can therefore have some error.

In [2] the authors demonstrated that there was mildly negative correlation between *Age* and *Bone mineral density* (BMD) (r=-0.381), and the association was highly significant (p=0.000). However, this thesis demonstrates that a large number of hypotheses are required to test the existence of a relation between two variables in order to ensure a complete assessment of dependencies. Moreover, when there are many sources of evidence a conclusion regarding a query requires all of them to be considered during an analysis.

Further, a multilevel regression model can only model one outcome variable but BNs are a suitable tool for modelling multiple outcomes. The methodology discussed in this thesis show that a BN can be used to understand various details of the problem domain.

8.3.2 Related approaches using Bayesian network modelling

A BN is a suitable tool to represent causal associations of a domain. But it is not always the case that a relation has must represent causation between two variables; instead a relation can simply be an association. Many BNs are learnt directly from data using algorithms that do not distinguish causation from associations [182][183][184][116]. But there are algorithms (such as Pearl's Inferred-Causation [41][185]) for discovering some causation from data when the direction of causal relations is known from some source other than data.

Cooper and Yoo [185] used a mixer of experimental and observational data for learning a causal Bayesian network. Their approach starts by constructing a hypothetical causal Bayesian network. They considered a set of assumptions such as the use of a) complete data without any missing data or hidden variable, b) all discrete variables, c) Dirchlet prior distribution of parameters and so on. Under these assumptions, a closed form Bayesian scoring metric is used to score causal networks that construct from experimental and observational data.

In Dekker et al.'s [7] learning approach, the domain experts make a draft structure at first. In [186], domain experts identify constraints on the model, and then the model is modified by means of a cost function which determines the model's differences from those constraints. Heckerman et al. [90] identify event equivalence and parameter modularity properties of metrics to simplify the use of expert knowledge, and apply a score metric and a search procedure to learn Bayesian networks from a combination of knowledge and statistical data.

While these studies have considered expert judgments along with data to learn BNs for a domain, we have considered ways to use data to assess expert-suggested causal relations in a BN. Suppose there is a variable O which is thought to be caused by three variables M, N and P. Using our methodology the domain expert can suggest that there are causal relations. We can further learn about the strength of these relations by doing Bayesian analysis. Although Bayesian analysis is used in algorithms for learning model

structures [187][188], Bayesian analysis to understand about the strength of relations in a Bayesian network is rare.

This thesis presents techniques to tackle the above issue. Moreover, one can now model causal BN using knowledge and prove that the relations of the BN are plausible against data with confidence. Next, rather than scenario analysis of multiple binary nodes, Bayesian analysis can reveal information about two ordered or categorical nodes at any time.

8.4 Summary

This chapter explains the methodology proposed separately from the case study. It shows how the techniques can be combined into a novel data analysis method, and discusses the development of a new tool and its features to apply the modelling method in any domain. Then the chapter shows that the use of multivariate regression model and P-values is inefficient to answer queries that arise from non-experimental data. In particular, results suggesting correlations and the strength of a relation are not sufficient for understanding causation which relies on the use of knowledge.

Chapter 9

Conclusion

This thesis has focused on the use of BNs for the analysis of evidence in health services derived from observational studies, currently considered the type of studies providing the weakest evidence. More specifically, it has proposed a novel way to use a BN model to combine the use of expert judgement and data from observational studies to answer queries.

The research hypotheses with which the thesis started are reviewed in Section 9.1, with a summary and evaluation of how the thesis has demonstrated the hypotheses. The way forward and possible directions for the future research are outlined in Section 9.2.

9.1 Review of the research hypotheses

9.1.1 Hypothesis 1: the need for new methods

It is important to propose a new method for analysing observational data and producing answer to queries.

The use of expert judgment is indispensible for analysing certain types of clinical queries as an expert can identify the relations that model a problem domain. Bayesian networks give the flexibility to combine data with knowledge, which we have exploited to generate statistical evidence in the context of an expert-derived BN model. This thesis focuses on using BNs for the problem of clinical evidence derived from observational studies.

It is well known that experimental trials are not a possible design option to generate evidence for all clinical questions of interest. Even when a trial is possible, conducting a trial is sometime impractical since the time and cost required are often very high. So observational studies are conducted instead and researchers have commonly used P-values and Confidence Intervals to assess the results produced. However, both the measures have been criticised for their improper use and inadequate interpretation of results.

Bayesian inference can overcome many limitations of the above measures of the classical inference method. This thesis showed that the key advantages that make the Bayesian inference method particularly suitable for evaluating the strength of evidence found from observational data are that it can:

- Report on the probability of interest for evaluation of the strength of evidence. According to critics, the probability of data given the null hypothesis is not essential measure of evidence. Bayesian methods of statistical inference let us calculate what really is required and that is the probability of a hypothesis given data.

- Help to measure evidence and to determine the strength of the evidence. This is done by the use of Bayes Factors. A Bayes Factor, which is the ratio of the probabilities of two competing hypotheses, enables one to decide the strength of the evidence based on its magnitude.

- Support adequate interpretation by quantifying uncertainty.

- Permit inductive inference to assess cause-effect relations. By giving the probability for a hypothesis on the basis of the data Bayesian methods permit inductive inference, which are more appropriate for assessing cause-effect relations

Conclusion

A Bayesian network model can represent many types of relations and let us to generate evidence for situations which require many inferences to be performed. In addition, BNs can:

- Model multiple outcomes and therefore, can be used to model complex health care settings such as MDT meetings and waiting times, and derive evidence for its better management.

- Represent both discrete and continuous variables within a framework.

- Provide higher accuracy in incorporating continuous variables using dynamic discretisation algorithms.

To understand the applications of BNs in the clinical and health service domain a survey was performed. This survey covered the types of application, the associated techniques of evaluation and their limitations. We found that:

- Many studies use train-test datasets for evaluation.

- The accuracy of predictions made by BNs has mostly been assessed using the states of one outcome.

- The performance of a BN has commonly measured using: area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value, and HL statistics.

- Models fit to data have been explained using P-values.

Overall, the survey showed that the existing techniques for using BNs in the clinical domain are not sufficient for analysing the strength of associations. This, together with the limitations of the measures of classical inference methods (Chapter 2) and the existing development methods of BNs (Chapters 3 and 6), confirms Hypothesis 1.

9.1.2 Hypotheses 2 & 3: The structure of a BN model from knowledge and data

Data, if available, can demonstrate existence of associations in an expert constructed Bayesian Network model.

Using both the knowledge of experts and data from an observational study we can form a BN to represent associations between its variables

This thesis used a case study and data collected from meetings of a Multidisciplinary Team (MDT) that treats patient suffering with cancer or suspected to have cancer to introduce our research techniques. Initially, it constructed a causal BN by considering an expert's assumptions about the existence and the direction of causal associations.

We evaluated the expert-judged causal associations in the initial BN against data and used the results to revise the initial BN model. The algorithmic steps followed for this are:

- Considered each causal fragment of the expert constructed BN.
- Considered the relations between the fragment variables as a hypothesis.
- Used the available data for testing the hypothesis against all competing hypotheses.
- Revealed if two variables that thought to have a causal relation between them receive support of association from the data.
- Constructed a BN with both knowledge and data.

Hence, Chapters 5 and 6 confirm Hypotheses 2 and 3.

9.1.3 Hypothesis 4: The strength of strong associations

For the BN model we can assess the strength of each association. The results from this assessment can then help to address a relevant query with confidence.

After evaluating the existence of each expert-derived relation, we then assess how strong this association is. The algorithmic steps followed for this are:

- Considered each strong association from the revised BN model.
- Determined the posterior distribution over the parameters of each strong association from data, using an auxiliary multinomial BN model.
- Performed hypothesis tests.
- Produced evidence about differences between the probabilities computed from the data.
- Demonstrated how these differences can be useful for decision-support.

Hence, Chapter 7 confirms Hypothesis 4.

9.1.4 Hypothesis 5: A method for analysing observational data

It is possible to apply the above techniques to successfully analyse observational data in any domain

Finally, the thesis presented a novel approach for analysing observational data by using knowledge and data. It:

- Combined the above evaluation techniques and presented the complete methodology for the analysis of data.

Conclusion

- Sowed that the approach is not restricted to the domain of our case study but can be applied to successfully analyse relations in any domain.

Hence, this part of the research (Chapter 8) confirms Hypothesis 5.

9.2 Future work

This research could be extended in the following ways:

- To extend the MDT model and collect more data for investigating how MDT meetings help patients after they receive treatments.

- Instead of keeping the direction of an expert's judged relation unchanged throughout, it would be interesting to investigate if there is a possible change of the direction.

- The structure leaning technique proposed in this study depends on using a uniform prior on each parameter. It would be useful to investigate the technique by using a non-uniform distribution based on some prior values to demonstrate that this will not provide any restriction to learn the existence of a fragment by following the method used in this thesis.

- Develop a single software tool from the studied methods for modelling BNs from knowledge and data, and analysing observational data.

- To speed up inference in BNs using more efficient inference algorithms.

- To guide users by providing an interface that allows sequential selection of the steps depicted in Figure 8-1.

- To ensure database connectivity – so that the users can complete each modelling step using the relevant data file.

Conclusion

- To automate calculation of joint probability of data and a Bayesian score – so that the users can find a complete structure without looking into detail.

Chapter 10

References

[1] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records," J Am Med Inform Assoc, vol. 16, no. 3, pp. 371–379, Jun. 2009.

[2] N. Sultan and J. Rao, "Association between periodontal disease and bone mineral density in postmenopausal women: a cross sectional study," Med Oral Patol Oral Cir Bucal, vol. 16, no. 3, pp. e440–447, May 2011.

[3] R. Sund, J. Riihimäki, M. Mäkelä, A. Vehtari, P. Lüthje, T. Huusko, and U. Häkkinen, "Modeling the length of the care episode after hip fracture: does the type of fracture matter?," Scand J Surg, vol. 98, no. 3, pp. 169–174, 2009.

[4] T. Charitos, L. Vandergaag, S. Visscher, K. Schurink, and P. Lucas, "A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients☆," Expert Systems with Applications, vol. 36, no. 2, pp. 1249–1258, Mar. 2009.

[5] M. Athanasiou and J. Y. Clark, "A Bayesian network model for the diagnosis of the caring procedure for wheelchair users with spinal injury," Comput Methods Programs Biomed, vol. 95, no. 2 Suppl, pp. S44–54, Aug. 2009.

[6] B. A. Ahmed, M. E. Matheny, P. L. Rice, J. R. Clarke, and O. I. Ogunyemi, "A comparison of methods for assessing penetrating trauma on retrospective multi-center data," J Biomed Inform, vol. 42, no. 2, pp. 308–316, Apr. 2009.

[7] A. Dekker, C. Dehing-Oberije, D. De Ruysscher, P. Lambin, A. Hope, K. Komati, G. Fung, S. Yu, W. De Neve, and Y. Lievens, "Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in

Handling Missing Data," in Machine Learning and Applications, 2009. ICMLA '09. International Conference on, 2009, pp. 494 –497.

[8] M. Donald, A. Cook, and K. Mengersen, "Bayesian network for risk of diarrhea associated with the use of recycled water," Risk Anal, vol. 29, no. 12, pp. 1672–1685, Dec. 2009.

[9] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous, "Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression," Hum. Reprod, vol. 21, no. 7, pp. 1824–1831, Jul. 2006.

[10] B. Biagioli, S. Scolletta, G. Cevenini, E. Barbini, P. Giomarelli, and P. Barbini, "A multivariate Bayesian model for assessing morbidity after coronary artery surgery," Crit Care, vol. 10, no. 3, p. R94, 2006.

[11] G. Marston and R. Watts, "Tampering with the evidence: a critical appraisal of evidence-based policy-making," The Drawing Board: an Australian Review of Public Affairs., vol. 3, no. 3, pp. 143–163, 2003.

[12] M. M. Boxer, S. K. Vinod, J. Shafiq, and K. J. Duggan, "Do multidisciplinary team meetings make a difference in the management of lung cancer?," Cancer, vol. 117, no. 22, pp. 5112–5120, Nov. 2011.

[13] F. C. Wright, C. De Vito, B. Langer, A. Hunter, and Expert Panel on Multidisciplinary Cancer Conference Standards, "Multidisciplinary cancer conferences: a systematic review and development of practice standards," Eur. J. Cancer, vol. 43, no. 6, pp. 1002–1010, Apr. 2007.

[14] D. G. Altman, "The scandal of poor medical research," BMJ, vol. 308, no. 6924, pp. 283–284, Jan. 1994.

[15] T. P. Shakespeare, V. J. Gebski, M. J. Veness, and J. Simes, "Improving interpretation of clinical studies by use of confidence levels, clinical significance

curves, and risk-benefit contours," The Lancet, vol. 357, no. 9265, pp. 1349–1353, Apr. 2001.

[16] S. J. Gilmore, "Evaluating statistics in clinical trials: making the unintelligible intelligible," Australas. J. Dermatol., vol. 49, no. 4, pp. 177–184; quiz 185–186, Nov. 2008.

[17] P. C. Gotzsche, "Believability of relative risks and odds ratios in abstracts: cross sectional study," BMJ, vol. 333, no. 7561, pp. 231–234, Jul. 2006.

[18] N. Houssami and R. Sainsbury, "Breast cancer: Multidisciplinary care and clinical outcomes," European Journal of Cancer, vol. 42, no. 15, pp. 2480–2491, Oct. 2006.

[19] M. Coory, P. Gkolia, I. A. Yang, R. V. Bowman, and K. M. Fong, "Systematic review of multidisciplinary teams in the management of lung cancer," Lung Cancer, vol. 60, no. 1, pp. 14–21, Apr. 2008.

[20] C.-Z. Du, J. Li, Y. Cai, Y.-S. Sun, W.-C. Xue, and J. Gu, "Effect of multidisciplinary team treatment on outcomes of patients with gastrointestinal malignancy," World J Gastroenterol, vol. 17, no. 15, pp. 2013–2018, Apr. 2011.

[21] J. H. Chang, E. Vines, H. Bertsch, D. L. Fraker, B. J. Czerniecki, E. F. Rosato, T. Lawton, E. F. Conant, S. G. Orel, L. Schuchter, K. R. Fox, N. Zieber, J. H. Glick, and L. J. Solin, "The impact of a multidisciplinary breast cancer center on recommendations for patient management: the University of Pennsylvania experience," Cancer, vol. 91, no. 7, pp. 1231–1237, Apr. 2001.

[22] L. M. Forrest, D. C. McMillan, C. S. McArdle, and D. J. Dunlop, "An evaluation of the impact of a multidisciplinary team, in a single centre, on treatment and survival in patients with inoperable non-small-cell lung cancer," Br J Cancer, vol. 93, no. 9, pp. 977–978, Oct. 2005.

[23] L. E. Horvath, E. Yordan, D. Malhotra, I. Leyva, K. Bortel, D. Schalk, P. Mellinger, M. Huml, C. Kesslering, and J. Huml, "Multidisciplinary Care in the Oncology Setting: Historical Perspective and Data From Lung and Gynecology Multidisciplinary Clinics," J Oncol Pract, vol. 6, no. 6, pp. e21–e26, Nov. 2010.

[24] I. J. Higginson, I. Finlay, D. M. Goodwin, A. M. Cook, K. Hood, A. G. K. Edwards, H.-R. Douglas, and C. E. Norman, “Do hospital-based palliative teams improve care for patients or families at the end of life?,” J Pain Symptom Manage, vol. 23, no. 2, pp. 96–106, Feb. 2002.

[25] J. Hearn and I. J. Higginson, “Do specialist palliative care teams improve outcomes for cancer patients? A systematic literature review,” Palliat Med, vol. 12, no. 5, pp. 317–332, Sep. 1998.

[26] A. Jansson, A. Isacsson, and L. H. Lindholm, “Organization of health care teams and the population’s contacts with primary care,” Scand J Prim Health Care, vol. 10, no. 4, pp. 257–265, Dec. 1992.

[27] R. Haward, Z. Amir, C. Borrill, J. Dawson, J. Scully, M. West, and R. Sainsbury, “Breast cancer teams: the impact of constitution, new cancer workload, and methods of operation on their effectiveness,” Br. J. Cancer, vol. 89, no. 1, pp. 15–22, Jul. 2003.

[28] Stroke Unit Trialists’ Collaboration, “Organised inpatient (stroke unit) care for stroke,” Cochrane Database Syst Rev, vol. 9, p. CD000197, 2013.

[29] V. Patkar, D. Acosta, T. Davidson, A. Jones, J. Fox, and M. Keshtgar, “Cancer Multidisciplinary Team Meetings: Evidence, Challenges, and the Role of Clinical Decision Support Technology,” International Journal of Breast Cancer, vol. 2011, p. e831605, Jul. 2011.

[30] P. V. Murray, M. E. R. O’Brien, R. Sayer, N. Cooke, G. Knowles, A. C. Miller, V. Varney, N. P. Rowell, A. R. Padhani, D. MacVicar, A. Norton, S. Ashley, and I. E. Smith, “The pathway study: results of a pilot feasibility study in patients suspected of having lung carcinoma investigated in a conventional chest clinic setting compared to a centralised two-stop pathway,” Lung Cancer, vol. 42, no. 3, pp. 283–290, Dec. 2003.

[31] A. G. Davison, C. D. Eraut, A. S. Haque, S. Doffman, A. Tanqueray, C. W. Trask, A. Lamont, R. Uppal, and A. Sharma, “Telemedicine for multidisciplinary lung cancer meetings,” J Telemed Telecare, vol. 10, no. 3, pp. 140–143, 2004.

[32] A. Fleissig, V. Jenkins, S. Catt, and L. Fallowfield, “Multidisciplinary teams in cancer care: are they effective in the UK?,” Lancet Oncol., vol. 7, no. 11, pp. 935–943, Nov. 2006.

[33] C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-spectral Methods,” Econometrica, vol. 37, no. 3, pp. 424–438, Aug. 1969.

[34] P. W. Holland, “Statistics and Causal Inference,” Journal of the American Statistical Association, vol. 81, no. 396, pp. 945–960, Dec. 1986.

[35] S. Eljamel, “Photodynamic applications in brain tumors: a comprehensive review of the literature,” Photodiagnosis Photodyn Ther, vol. 7, no. 2, pp. 76–85, Jun. 2010.

[36] C. R. UK, “Biological therapies for chronic myeloid leukaemia (CML),” 22-Aug-2011. [Online]. Available: http://cancerhelp.cancerresearchuk.org/type/cml/treatment/biological-therapies-for-chronic-myeloid-leukaemia. [Accessed: 06-May-2012].

[37] P. M. Ho, P. N. Peterson, and F. A. Masoudi, “Evaluating the Evidence Is There a Rigid Hierarchy?,” Circulation, vol. 118, no. 16, pp. 1675–1684, Oct. 2008.

[38] V. Lutje, A. Gerritsen, and N. Siegfried, “Randomized controlled trials of malaria intervention trials in Africa, 1948 to 2007: a descriptive analysis,” Malar J, vol. 10, p. 61, Mar. 2011.

[39] D. A. Berry, “Interim analyses in clinical trials: Classical vs. bayesian approaches,” Statistics in Medicine, vol. 4, no. 4, pp. 521–526, Oct. 1985.

[40] M. A. Hernán and J. M. Robins, “Instruments for causal inference: an epidemiologist’s dream?,” Epidemiology, vol. 17, no. 4, pp. 360–372, Jul. 2006.

[41] J. Pearl, Causality: Models, Reasoning, and Inference, 1st Edition. Cambridge University Press, 2000.

[42] D. N. Wijeysundera, P. C. Austin, J. E. Hux, W. S. Beattie, and A. Laupacis, “Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials,” J. Clin. Epidemiol., vol. 62, no. 1, pp. 13–21, Jan. 2009.

[43] P. Armitage, G. Berry, and J. N. S. Matthews, Statistical Methods in Medical Research, 2nd Edition. Oxford:Blackwell Scientific, 1994.

[44] J. Fethney, “Statistical and clinical significance, and how to use confidence intervals to help interpret both,” Australian Critical Care, vol. 23, no. 2, pp. 93–97, May 2010.

[45] A. K. Akobeng, “Confidence intervals and p-values in clinical decision making,” Acta Paediatrica, vol. 97, no. 8, pp. 1004–1007, Aug. 2008.

[46] J. A. C. Sterne, “Sifting the evidence---what’s wrong with significance tests? Another comment on the role of statistical methods,” BMJ, vol. 322, no. 7280, pp. 226–231, Jan. 2001.

[47] “What Is the Chance That This Study Is Clinically Significant?” [Online]. Available: http://www.acponline.org/clinical_information/journals_publications/ecp/sepoct99/froehlich.htm#11. [Accessed: 18-Mar-2012].

[48] “Balancing Statistical and Clinical Significance in Evaluating Treatment Effects,” Postgrad Med J, vol. 77, no. 905, pp. 201–204, Mar. 2001.

[49] L. E. Braitman, “Confidence intervals assess both clinical significance and statistical significance,” Ann. Intern. Med., vol. 114, no. 6, pp. 515–517, Mar. 1991.

[50] D. G. Altman and M. J. Gardner, “Confidence intervals for research findings,” BJOG: An International Journal of Obstetrics & Gynaecology, vol. 99, no. 2, pp. 90–91, Feb. 1992.

[51] J. E. Harrison, “Evidence-Based Orthodontics—How Do I Assess the Evidence?,” J. Orthod., vol. 27, no. 2, pp. 189–197, Jun. 2000.

[52] K. B. Y. Chan, M. Man-Son-Hing, F. J. Molnar, and A. Laupacis, “How well is the clinical importance of study results reported? An assessment of randomized controlled trials,” CMAJ, vol. 165, no. 9, pp. 1197–1202, Oct. 2001.

[53] G. Steven, “A Dirty Dozen: Twelve P-Value Misconceptions,” Seminars in Hematology, vol. 45, no. 3, pp. 135–140, Jul. 2008.

[54] N. E. Fenton and M. Neil, Risk assessment and decision analysis with Bayesian networks. Boca Raton: Taylor & Francis, 2012.

[55] S. T. Ziliak and D. N. McCloskey, The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor: University of Michigan Press, 2008.

[56] S. N. Goodman, “Introduction to Bayesian methods I: measuring the strength of evidence,” Clin Trials, vol. 2, no. 4, pp. 282–290; discussion 301–304, 364–378, 2005.

[57] J. M. Brophy and L. Joseph, “Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes,” JAMA, vol. 273, no. 11, pp. 871–875, Mar. 1995.

[58] S. N. Goodman, “Toward evidence-based medical statistics. 2: The Bayes factor,” Ann. Intern. Med., vol. 130, no. 12, pp. 1005–1013, Jun. 1999.

[59] A. R. Feinstein, “P-Values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin,” Journal of Clinical Epidemiology, vol. 51, no. 4, pp. 355–360, Apr. 1998.

[60] J. J. Lee, “Demystify Statistical Significance—Time to Move on From the P Value to Bayesian Analysis,” JNCI J Natl Cancer Inst, vol. 103, no. 1, pp. 2–3, Jan. 2011.

[61] S. N. Goodman, “Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy,” Ann Intern Med, vol. 130, no. 12, pp. 995–1004, Jun. 1999.

[62] D. N. Kyriacou, “Evidence-based medical decision making: deductive versus inductive logical thinking,” Acad Emerg Med, vol. 11, no. 6, pp. 670–671, Jun. 2004.

[63] F. Davidoff, “Standing Statistics Right Side Up,” Ann Intern Med, vol. 130, no. 12, pp. 1019–1021, Jun. 1999.

[64] P. R. Burton, “Helping doctors to draw appropriate inferences from the analysis of medical studies,” Statistics in Medicine, vol. 13, no. 17, pp. 1699–1713, Oct. 2006.

[65] T. Fahey, N. Stocks, and T. Thomas, “Quantitative systematic review of randomised controlled trials comparing antibiotic with placebo for acute cough in adults,” BMJ, vol. 316, no. 7135, pp. 906–910, Mar. 1998.

[66] R. E. Kass and A. E. Raftery, Bayes Factors. 1995.

[67] W. S. Browner and T. B. Newman, “Are All Significant P Values Created Equal? The Analogy Between Diagnostic Tests and Clinical Research,” JAMA: The Journal of the American Medical Association, vol. 257, no. 18, pp. 2459–2463, May 1987.

[68] G. H. Skrepnek, “The Contrast and Convergence of Bayesian and Frequentist Statistical Approaches in Pharmacoeconomic Analysis,” PharmacoEconomics, vol. 25, no. 8, pp. 649–664, Jul. 2007.

[69] D. Ashby and A. F. Smith, “Evidence-based medicine as Bayesian decision-making,” Stat Med, vol. 19, no. 23, pp. 3291–3305, Dec. 2000.

[70] S. N. Goodman, “Of P-values and Bayes: a modest proposal,” Epidemiology, vol. 12, no. 3, pp. 295–297, May 2001.

[71] “Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests,” Perspectives on Psychological Science, vol. 6, pp. 291–298, 2011.

[72] H. Jeffreys, Theory of Probability. Oxford University Press, 1998.

[73] T. D. Nielsen and F. V. JENSEN, Bayesian Networks and Decision Graphs, Softcover reprint of hardcover 2nd ed. 2007. Springer, 2010.

[74] S. Lauritzen and D. Spiegelhalter, “Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 50, no. 2, 1988.

[75] M. Neil, M. Tailor, D. Marquez, N. Fenton, and P. Hearty, “Modelling dependable systems using hybrid Bayesian networks,” Reliability Engineering & System Safety, vol. 93, no. 7, pp. 933–939, Jul. 2008.

[76] C. A. M. Schurink, P. J. F. Lucas, I. M. Hoepelman, and M. J. M. Bonten, “Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units,” Lancet Infect Dis, vol. 5, no. 5, pp. 305–312, May 2005.

[77] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Calif.: Morgan Kaufmann Publishers, 1997.

[78] F. Jensen, An Introduction to Bayesian Networks. UCL Press, 1996.

[79] M. Neil, M. Tailor, and D. Marquez, “Inference in hybrid Bayesian networks using dynamic discretization,” Statistics and Computing, vol. 17, 2007.

[80] D. L. Keefer and S. E. Bodily, “Three-Point Approximations for Continuous Random Variables,” Management Science, vol. 29, no. 5, pp. 595–609, May 1983.

[81] R. T. Clemen, Making Hard Decisions: An Introduction to Decision Analysis. Duxbury Press, 1996.

[82] D. Marquez, M. Neil, and N. Fenton, “Improved reliability modeling using Bayesian networks and dynamic discretization,” Reliability Engineering System Safety, vol. 95, no. 4, pp. 412–425, 2010.

[83] N. Fenton, L. Radlinski, and M. Neil, Improved Bayesian Networks for Software Project Risk Assessment Using Dynamic Discretisation. .

[84] A. V. Kozlov and D. Koller, “Nonuniform dynamic discretization in hybrid networks,” Association for Uncertainty in Artificial Intelligence, pp. 314–325, 1997.

[85] M. N. N. Fenton and Q. Mary, Using Bayesian Networks and Simulation for Data Fusion and Risk Analysis. .

[86] J. Pearl, Fusion, Propagation, and Structuring in Belief Networks. 1986.

[87] R. D. Shachter, “Probabilistic Inference and Influence Diagrams,” Operations Research, vol. 36, no. 4, pp. 589–604, Jul. 1988.

[88] N. Zhang and D. Poole, “A simple approach to Bayesian network computations,” presented at the Proceedings of the Tenth Canadian Conference on Artificial Intelligence, 1994, pp. 171–178.

[89] P. Lucas, “Knowledge acquisition for decision-theoretic expert systems,” AISB Quarterly, vol. 94, pp. 23–33, 1996.

[90] D. Heckerman, “A Tutorial on Learning With Bayesian Networks,” Learning in Graphical Models, 1996.

[91] M. Neil, N. Fenton, and L. Nielsen, Building Large-Scale Bayesian Networks. 1999.

[92] M. Verduijn, N. Peek, P. M. J. Rosseel, E. de Jonge, and B. A. J. M. de Mol, “Prognostic Bayesian networks: I: Rationale, learning procedure, and clinical use,” Journal of Biomedical Informatics, vol. 40, no. 6, pp. 609–618, Dec. 2007.

[93] K. B. Laskey and S. M. Mahoney, “Network Fragments: Representing Knowledge for Constructing Probabilistic Models,” pp. 334–341, 1997.

[94] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, “Toward normative expert systems: Part I. The Pathfinder project,” Methods Inf Med, vol. 31, no. 2, pp. 90–105, Jun. 1992.

[95] K. G. Olesen, U. Kjaerulff, F. Jensen, F. V. Jensen, B. Falck, S. Andreassen, and S. K. Andersen, “A munin network for the median nerve-a case study on loops,” Appl. Artif. Intell., vol. 3, no. 2–3, pp. 385–403, Oct. 1989.

[96] N. Hoot and D. Aronsky, “Using Bayesian networks to predict survival of liver transplant patients,” AMIA Annu Symp Proc, pp. 345–349, 2005.

[97] S. Sadeghi, A. Barzi, N. Sadeghi, and B. King, “A Bayesian model for triage decision support,” International Journal of Medical Informatics, vol. 75, no. 5, pp. 403–411, May 2006.

[98] E. S. Burnside, D. L. Rubin, J. P. Fine, R. D. Shachter, G. A. Sisney, and W. K. Leung, “Bayesian Network to Predict Breast Cancer Risk of Mammographic Microcalcifications and Reduce Number of Benign Biopsy Results: Initial Experience1,” Radiology, vol. 240, no. 3, pp. 666–673, 2006.

[99] C. E. Kahn, L. M. Roberts, K. Wang, D. Jenks, and P. Haddawy, “Preliminary investigation of a Bayesian network for mammographic diagnosis of breast cancer.,” Proc Annu Symp Comput Appl Med Care, pp. 208–212, 1995.

[100] I. J. Good, “The Estimation of Probabilities: An Essay on Modern Bayesian Methods,” Mar. 2003.

[101] D. Jenkinson, The Elicitation of Probabilities- A Review of the Statistical Literature. 2005.

[102] P. E. Johnson, A. S. Duran, F. Hassebrock, J. Moller, M. Prietula, P. J. Feltovich, and D. B. Swanson, “Expertise and Error in Diagnostic Reasoning*,” Cognitive Science, vol. 5, no. 3, pp. 235–283, 1981.

[103] E. W. Watt, E. Watt, A. A. T. Bui, and A. A. Bui, “Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative,” AMIA Annu Symp Proc, pp. 788–792, 2008.

[104] D. M. Berwick, H. V. Fineberg, and M. C. Weinstein, “When doctors meet numbers,” Am. J. Med., vol. 71, no. 6, pp. 991–998, Dec. 1981.

[105] K. Miettinen and M. Juhola, “Classification of Otoneurological Cases According to Bayesian Probabilistic Models,” J. Med. Syst., vol. 34, no. 2, pp. 119–130, Apr. 2010.

[106] Wojtek Michalowski, Szymon Wilk, Anthony Thijssen, and Mingmei Li, “Using a Bayesian belief network model to categorize length of stay for radical prostatectomy patients,” Health Care Management Science, vol. 9, no. 4, pp. 341–348, 2006.

[107] S. Suebnukarn, N. Rungcharoenporn, and S. Sangsuratham, “A Bayesian decision support model for assessment of endodontic treatment outcome,” Oral Surg Oral Med Oral Pathol Oral Radiol Endod, vol. 106, no. 3, pp. e48–58, Sep. 2008.

[108] G. F. Cooper and T. Dietterich, “A Bayesian method for the induction of probabilistic networks from data,” in Machine Learning, 1992, pp. 309–347.

[109] R. Daly, Q. Shen, and S. Aitken, “Learning Bayesian networks: approaches and issues,” The Knowledge Engineering Review, vol. 26, no. 02, pp. 99–157, May 2011.

[110] W. Michalowski, S. Wilk, A. Thijssen, and M. Li, “Using a Bayesian belief network model to categorize length of stay for radical prostatectomy patients,” Health Care Manag Sci, vol. 9, no. 4, pp. 341–348, Nov. 2006.

[111] S. Suebnukarn, N. Rungcharoenporn, and S. Sangsuratham, “A Bayesian decision support model for assessment of endodontic treatment outcome,” Oral Surg Oral Med Oral Pathol Oral Radiol Endod, vol. 106, no. 3, pp. e48–58, Sep. 2008.

[112] A. L. Madsen, M. Lang, U. B. Kjærulff, and F. Jensen, “The Hugin Tool for Learning Bayesian Networks,” in In Proceedings of 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 2003, pp. 594–605.

[113] M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell, “Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology,” Am. J. Epidemiol., vol. 155, no. 2, pp. 176–184, Jan. 2002.

[114] S. Greenland, J. Pearl, and J. M. Robins, “Causal diagrams for epidemiologic research,” Epidemiology, vol. 10, no. 1, pp. 37–48, Jan. 1999.

[115] J. Agosta, “Automated Refinement of Bayes Networks’ Parameters based on Test Ordering Constraints,” presented at the Advances in Neural Information Processing Systems 24, 2011.

[116] D. Eaton, “Bayesian structure learning using dynamic programming and MCMC,” in In UAI, 2007b.

[117] A. J. Hartemink, “Banjo: Bayesian Network Inference with Java Objects,” 15-Oct-2012. [Online]. Available: http://www.cs.duke.edu/~amink/software/banjo/.

[118] Roberts, “DBNbox,” 15-Oct-2012. [Online]. Available: http://www.robots.ox.ac.uk/~parg/software.html.

[119] D. J. Wilkinson, GDAGsim: Sparse matrix algorithms for Bayesian computation. .

[120] F. G. Cozman, “JavaBayes Version 0.347 Bayesian Networks in Java,” 15-Oct-2012. [Online]. Available: http://sites.poli.usp.br/pmr/ltd/Software/javabayes/Home/.

[121] C. D. Susanne G. Bttcher, “Learning Bayesian Networks with R,” 2003.

[122] N. Friedman and G. Elidan, “LibB for Windows/Linux Programs,” 15-Oct-2012. [Online]. Available: http://www.cs.huji.ac.il/labs/compbio/LibB/.

[123] Shenoy and Shafer, “Pulcinella,” 15-Oct-2012. [Online]. Available: http://iridia.ulb.ac.be/pulcinella/.

[124] D. J. Spiegelhalter, A. Thomas, and N. Best, “The BUGS Project winbugs,” 15-Oct-2012. [Online]. Available: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

[125] M. Ladeira and R. N. Carvalho, “UnBBayes,” 15-Oct-2012. [Online]. Available: http://sourceforge.net/projects/unbbayes/.

[126] B. Winn, “Vibes: Variational inference for bayesian networks.,” 15-Oct-2012. [Online]. Available: http://vibes.sourceforge.net/.

[127] “Agenarisk software,” 15-Oct-2012. [Online]. Available: http://www.agenarisk.com/.

[128] “BayesiaLab,” 15-Oct-2012. [Online]. Available: http://www.bayesia.com/en/products/bayesialab.php.

[129] “Bayesian Knowledge Discovery,” 15-Oct-2012. [Online]. Available: http://projects.kmi.open.ac.uk/bkd/.

[130] “Lumina decision systems,” 15-Oct-2012. [Online]. Available: http://www.lumina.com/.

[131] “Netica,” 15-Oct-2012. [Online]. Available: http://www.norsys.com/.

[132] K. Murphy, “Software Packages for Graphical Models / Bayesian Networks.” [Online]. Available: http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html.

[133] S. M. Alvarez, B. A. Poelstra, and R. S. Burd, “Evaluation of a Bayesian decision network for diagnosing pyloric stenosis,” Journal of Pediatric Surgery, vol. 41, no. 1, pp. 155–161, Jan. 2006.

[134] J. A. Kline, A. J. Novobilski, C. Kabrhel, P. B. Richman, and D. M. Courtney, “Derivation and Validation of a Bayesian Network to Predict Pretest Probability of Venous Thromboembolism,” Annals of Emergency Medicine, vol. 45, no. 3, pp. 282–290, Mar. 2005.

[135] D. Luciani, S. Cavuto, L. Antiga, M. Miniati, S. Monti, M. Pistolesi, and G. Bertolini, “Bayes pulmonary embolism assisted diagnosis: a new expert system for clinical use,” Emerg Med J, vol. 24, no. 3, pp. 157–164, Mar. 2007.

[136] P. Haddawy, C. E. Kahn Jr, and M. Butarbutar, “A Bayesian network model for radiological diagnosis and procedure selection: work-up of suspected gallbladder disease,” Med Phys, vol. 21, no. 7, pp. 1185–1192, Jul. 1994.

[137] N. Cruz-Ramírez, H. G. Acosta-Mesa, H. Carrillo-Calvet, L. A. Nava-Fernández, and R. E. Barrientos-Martínez, “Diagnosis of breast cancer using Bayesian networks: a case study,” Comput. Biol. Med, vol. 37, no. 11, pp. 1553–1564, Nov. 2007.

[138] J. A. Kline, A. J. Novobilski, C. Kabrhel, P. B. Richman, and D. M. Courtney, “Derivation and validation of a Bayesian network to predict pretest probability of venous thromboembolism,” Ann Emerg Med, vol. 45, no. 3, pp. 282–290, Mar. 2005.

[139] D. L. Sanders and D. Aronsky, “Prospective evaluation of a Bayesian Network for detecting asthma exacerbations in a Pediatric Emergency Department,” AMIA Annu Symp Proc, p. 1085, 2006.

[140] B. E. Himes, A. C. Wu, Q. L. Duan, B. Klanderman, A. A. Litonjua, K. Tantisira, M. F. Ramoni, and S. T. Weiss, “Predicting response to short-acting bronchodilator medication using Bayesian networks,” Pharmacogenomics, vol. 10, no. 9, pp. 1393–1412, Sep. 2009.

[141] D. L. Sanders and D. Aronsky, “Detecting asthma exacerbations in a pediatric emergency department using a Bayesian network,” AMIA Annu Symp Proc, pp. 684–688, 2006.

[142] D. Aronsky and P. J. Haug, “Automatic identification of patients eligible for a pneumonia guideline.,” Proc AMIA Symp, pp. 12–16, 2000.

[143] Y. I. Liu, A. Kamaya, T. S. Desser, and D. L. Rubin, “A bayesian network for differentiating benign from malignant thyroid nodules using sonographic and demographic features,” AJR Am J Roentgenol, vol. 196, no. 5, pp. W598–605, May 2011.

[144] S. Chattopadhyay, R. M. Davis, D. D. Menezes, G. Singh, R. U. Acharya, and T. Tamura, “Application of Bayesian Classifier for the Diagnosis of Dental Pain,” J Med Syst, vol. 36, no. 3, pp. 1425–1439, Jun. 2012.

[145] Y. Lee, N. Kim, K.-S. Cho, S.-H. Kang, D. Y. Kim, Y. Y. Jung, and J. K. Kim, “Bayesian classifier for predicting malignant renal cysts on MDCT: early clinical experience,” AJR Am J Roentgenol, vol. 193, no. 2, pp. W106–111, Aug. 2009.

[146] J. Wu, G. Zhang, and Y. Cao, “Research on Cerebral Aneurysm Image Recognition Method Using Bayesian Classification,” 2009.

[147] E. S. Burnside, “Bayesian networks: computer-assisted diagnosis support in radiology,” Acad Radiol, vol. 12, no. 4, pp. 422–430, Apr. 2005.

[148] R. Blanco, I. Inza, M. Merino, J. Quiroga, and P. Larrañaga, “Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS,” J Biomed Inform, vol. 38, no. 5, pp. 376–388, Oct. 2005.

[149] A. Stojadinovic, C. Eberhardt, L. Henry, J. Eberhardt, E. A. Elster, G. E. Peoples, A. Nissan, and C. D. Shriver, “Development of a Bayesian classifier for breast cancer risk stratification: a feasibility study,” Eplasty, vol. 10, p. e25, 2010.

[150] M. Verduijn, P. M. J. Rosseel, N. Peek, E. de Jonge, and B. A. J. M. de Mol, “Prognostic Bayesian networks II: an application in the domain of cardiac surgery,” J Biomed Inform, vol. 40, no. 6, pp. 619–630, Dec. 2007.

[151] R. S. Burd, M. Ouyang, and D. Madigan, “Bayesian Logistic Injury Severity Score: A Method for Predicting Mortality Using International Classification of Disease-9 Codes,” Academic Emergency Medicine, vol. 15, no. 5, pp. 466–475, 2008.

[152] P. Sebastiani, V. G. Nolan, C. T. Baldwin, M. M. Abad-Grau, L. Wang, A. H. Adewoye, L. C. McMahon, L. A. Farrer, J. G. Taylor, G. J. Kato, M. T. Gladwin, and M. H. Steinberg, “A network model to predict the risk of death in sickle cell disease,” Blood, vol. 110, no. 7, pp. 2727–2735, Oct. 2007.

[153] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, and A. L. A. J. Dekker, “Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy,” Med Phys, vol. 37, no. 4, pp. 1401–1407, Apr. 2010.

[154] C. Berzuini, R. Bellazzi, S. Quaglini, and D. J. Spiegelhalter, “Bayesian networks for patient monitoring,” Artificial Intelligence in Medicine, vol. 4, no. 3, pp. 243–260, May 1992.

[155] A. H. Marshall, S. I. McClean, C. M. Shapcott, I. R. Hastie, and P. H. Millard, “Developing a Bayesian belief network for the management of geriatric hospital care,” Health Care Management Science, vol. 4, no. 1, pp. 25–30, Feb. 2001.

[156] K. Miettinen and M. Juhola, “Classification of Otoneurological Cases According to Bayesian Probabilistic Models,” J Med Syst, vol. 34, no. 2, pp. 119–130, Oct. 2008.

[157] G. Reynolds, A. Peet, and T. Arvanitis, “Generating prior probabilities for classifiers of brain tumours using belief networks,” BMC Medical Informatics and Decision Making, vol. 7, no. 1, p. 27, 2007.

[158] S. A. Vinterbo, S. A. Vinterbo, N. Trondheim, and N. T. Universitet, “Predictive Models in Medicine: Some Methods for Construction and Adaptation,” 1999.

[159] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression. John Wiley & Sons, 2000.

[160] S. G. Pauker and J. P. Kassirer, “The threshold approach to clinical decision making,” N. Engl. J. Med., vol. 302, no. 20, pp. 1109–1117, May 1980.

[161] S. M. Maskery, H. Hu, J. Hooke, C. D. Shriver, and M. N. Liebman, “A Bayesian derived network of breast pathology co-occurrence,” Journal of Biomedical Informatics, vol. 41, no. 2, pp. 242–250, Apr. 2008.

[162] L. Ohno-Machado and S. Vinterbo, “Effects of case removal in prognostic models,” Methods Inf Med, vol. 40, no. 1, pp. 32–38, Mar. 2001.

[163] S. A. R. Nouraei, Q. J. M. Huys, P. Chatrath, J. Powles, and J. P. Harcourt, “Screening patients with sensorineural hearing loss for vestibular schwannoma using a Bayesian classifier,” Clin Otolaryngol, vol. 32, no. 4, pp. 248–254, Aug. 2007.

[164] K. Johansen, M. Daines, T. Howey, D. Helfet, and S. T. Hansen Jr, “Objective criteria accurately predict amputation following lower extremity trauma,” J Trauma, vol. 30, no. 5, pp. 568–572; discussion 572–573, May 1990.

[165] B. Yet, Z. Perkins, W. Marsh, and N. Fenton, “Towards a Method of Building Causal Bayesian Networks for Prognostic Decision Support.”

[166] V. Oluwole Ogunsanya, “Decision Support using Bayesin Networks for Clinical Decision Making.” Aug-2011.

[167] M. Birchall, D. Bailey, and P. King, “Effect of process standards on survival of patients with head and neck cancer in the south and west of England,” Br. J. Cancer, vol. 91, no. 8, pp. 1477–1481, Oct. 2004.

[168] M. Gabel, N. E. Hilton, and S. D. Nathanson, “Multidisciplinary breast cancer clinics. Do they work?,” Cancer, vol. 79, no. 12, pp. 2380–2384, Jun. 1997.

[169] J. H. Chang, E. Vines, H. Bertsch, D. L. Fraker, B. J. Czerniecki, E. F. Rosato, T. Lawton, E. F. Conant, S. G. Orel, L. Schuchter, K. R. Fox, N. Zieber, J. H. Glick, and L. J. Solin, “The impact of a multidisciplinary breast cancer center on recommendations for patient management,” Cancer, vol. 91, no. 7, pp. 1231–1237, Apr. 2001.

[170] N. Friedman, M. Goldszmidt, and A. Wyner, “Data analysis with bayesian networks: a bootstrap approach,” in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, San Francisco, CA, USA, 1999, pp. 196–205.

[171] B. Efron and R. Tibshirani, An Introduction to the Bootstrap. Chapman & Hall, 1993.

[172] D. Heckerman, C. Meek, and G. Cooper, “A Bayesian Approach to Causal Discovery,” 1997.

[173] S. Burton, G. Brown, I. R. Daniels, A. R. Norman, B. Mason, and D. Cunningham, “MRI directed multidisciplinary team preoperative treatment strategy: the way to eliminate positive circumferential margins?,” Br. J. Cancer, vol. 94, no. 3, pp. 351–357, Feb. 2006.

[174] M. M. Boxer, S. K. Vinod, J. Shafiq, and K. J. Duggan, “Do multidisciplinary team meetings make a difference in the management of lung cancer?,” Cancer, vol. 117, no. 22, pp. 5112–5120, 2011.

[175] A. Caudron, G. Chaby, A. Dadban, C. Andrejak, F. Dhaille, M. Bagot, and C. Lok, “Multidisciplinary team meetings in Oncology: first analysis of benefits and evaluation of activity in a Dermatology unit in France,” Eur J Dermatol, vol. 20, no. 6, pp. 778–784, Dec. 2010.

[176] R. Haining, G. Li, R. Maheswaran, M. Blangiardo, J. Law, N. Best, and S. Richardson, “Inference from ecological models: Estimating the relative risk of stroke from air pollution exposure using small area data,” Spatial and Spatio-temporal Epidemiology, vol. 1, no. 2–3, pp. 123–131, Jul. 2010.

[177] A. C. CONSTANTINOU, N. E. FENTON, and M. NEIL, “Profiting from an Inefficient Association Football Gambling Market: Prediction, Risk and Uncertainty Using Bayesian Networks,” Under Review. Draft available at: http://www. constantinou. info/downloads/papers/pi-model12. pdf, 2012.

[178] Y. Zhou, N. Fenton, and M. Neil, “Bayesian Network Approach to Multinomial Parameter Learning using Data and Expert Judgements.,” Journal of Approximate Reasoning, 2013.

[179] On Regression Analysis Of The Relationship Between Age And Blood Cholesterol On Blood Pressure pdf free ebook download. .

[180] J. F. Tomera and C. Harakal, “Multiple linear regression analysis of blood pressure, hypertrophy, calcium and cadmium in hypertensive and non-hypertensive states,” Food Chem. Toxicol., vol. 35, no. 7, pp. 713–718, Jul. 1997.

[181] R. J. Rona, S. Qureshi, and S. Chinn, “Factors related to total cholesterol and blood pressure in British 9 year olds,” J Epidemiol Community Health, vol. 50, no. 5, pp. 512–518, Oct. 1996.

[182] J. Cussens, “Bayesian network learning with cutting planes,” presented at the Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI-11), 2011, pp. 153–160.

[183] T. Jaakkola, D. Sontag, A. Globerson, and M. Meila, “Learning bayesian network structure using lp relaxations,” MIT web domain, May 2010.

[184] W. Buntine, A Guide to the Literature on Learning Probabilistic Networks From Data. 1996.

[185] G. F. Cooper and C. Yoo, “Causal Discovery from a Mixture of Experimental and Observational Data,” arXiv:1301.6686 [cs], Jan. 2013.

[186] W. Wiegerinck, Modeling Bayesian Networks by Learning from Experts. .

[187] M. J. Price, N. J. Welton, and A. E. Ades, “Parameterization of treatment effects for meta-analysis in multi-state Markov models,” Stat Med, vol. 30, no. 2, pp. 140–151, Jan. 2011.

[188] S. Visweswaran, D. C. Angus, M. Hsieh, L. Weissfeld, D. Yealy, and G. F. Cooper, “Learning patient-specific predictive models from clinical data,” J Biomed Inform, vol. 43, no. 5, pp. 669–685, Oct. 2010.

Appendix A

Establishing plausible causal relations

A.1 Plausible relations - number of meetings, year and diagnosis

Table A-1 The posterior probability of each data point for each learned parameter of H2: Number of meetings is dependent on Diagnosis and Year

Year	Diagnosis	P(data point | Number of meetings =			Year	Diagnosis	P(data point| Number of meetings =		
		1	2	3			1	2	3
2005	BP	0.099	0.099	0.505	2007	ML	0.050	0.053	0.124
	BL	0.119	0.119	0.506		MGB	0.089	0.090	0.254
	BGB	0.131	0.131	0.509		Multiple	0.214	0.272	0.272
	MP	0.069	0.070	0.251		Unknown	0.105	0.113	0.195
	ML	0.059	0.061	0.139	2008	BP	0.078	0.079	0.191
	MGB	0.091	0.093	0.255		BL	0.099	0.099	0.505
	Multiple	0.231	0.231	0.538		BGB	0.126	0.126	0.261
	Unknown	0.109	0.109	0.505		MP	0.065	0.067	0.189
2006	BP	0.176	0.176	0.514		ML	0.055	0.056	0.138
	BL	0.205	0.266	0.515		MGB	0.081	0.082	0.190
	BGB	0.157	0.157	0.517		Multiple	0.151	0.151	0.514
	MP	0.107	0.118	0.195		Unknown	0.067	0.070	0.159
	ML	0.088	0.097	0.161	2009	BP	0.271	0.521	0.272
	MGB	0.174	0.174	0.512		BL	0.231	0.244	0.292
	Multiple	0.571	0.570	0.570		BGB	0.217	0.217	0.523
	Unknown	0.200	0.200	0.510		MP	0.151	0.151	0.514
2007	BP	0.104	0.105	0.506		ML	0.097	0.098	0.506
	BL	0.145	0.156	0.261		MGB	0.292	0.243	0.292

	BGB	0.152	0.152	0.509			Multiple	0.666	0.666	0.666
	MP	0.064	0.067	0.158			Unknown	0.135	0.135	0.510

Table A-2 The posterior probability of each data point for each learned parameter of H3: Number of meetings is dependent on Year

Year	Diagnosis	P(data point | Number of meetings =				Year	Diagnosis	P(data point| Number of meetings =		
		1	2	3				1	2	3
2005	BP	0.124	0.112	0.512		2007	ML	0.041	0.054	0.151
	BL	0.084	0.103	0.540			MGB	0.108	0.092	0.319
	BGB	0.152	0.165	0.635			Multiple	0.189	0.136	0.267
	MP	0.063	0.068	0.302			Unknown	0.143	0.147	0.208
	ML	0.037	0.058	0.077		2008	BP	0.102	0.105	0.256
	MGB	0.043	0.050	0.333			BL	0.062	0.095	0.271
	Multiple	0.189	0.171	0.890			BGB	0.090	0.104	0.340
	Unknown	0.037	0.051	0.453			MP	0.036	0.045	0.221
2006	BP	0.228	0.236	0.593			ML	0.047	0.051	0.171
	BL	0.204	0.167	0.612			MGB	0.074	0.081	0.248
	BGB	0.036	0.019	0.652			Multiple	0.183	0.202	0.584
	MP	0.120	0.147	0.195			Unknown	0.088	0.085	0.197
	ML	0.105	0.118	0.152		2009	BP	0.038	0.014	0.262
	MGB	0.205	0.230	0.558			BL	0.261	0.319	0.179
	Multiple	0.482	0.547	0.903			BGB	0.138	0.188	0.690
	Unknown	0.102	0.154	0.509			MP	0.167	0.192	0.562
2007	BP	0.135	0.136	0.296			ML	0.033	0.039	0.331
	BL	0.182	0.185	0.341			MGB	0.035	0.090	0.179
	BGB	0.075	0.126	0.443			Multiple	0.613	0.660	0.962
	MP	0.079	0.083	0.203			Unknown	0.117	0.154	0.477

Table A-3 The posterior probability of each data point for each learned parameter of H4: Number of meetings is dependent on Diagnosis

Year	Diagnosis	P(data point | Number of meetings =				Year	Diagnosis	P(data point| Number of meetings =		
		1	2	3				1	2	3
2005	BP	0.120	0.110	0.428		2007	ML	0.057	0.052	0.154
	BL	0.132	0.145	0.463			MGB	0.092	0.100	0.297
	BGB	0.136	0.144	0.623			Multiple	0.212	0.162	0.279
	MP	0.078	0.086	0.220			Unknown	0.131	0.146	0.177
	ML	0.075	0.077	0.177		2008	BP	0.058	0.071	0.206
	MGB	0.109	0.107	0.318			BL	0.113	0.097	0.403
	Multiple	0.164	0.140	0.742			BGB	0.048	0.074	0.246
	Unknown	0.061	0.095	0.293			MP	0.079	0.078	0.233

2006	BP	0.136	0.158	0.670			ML	0.046	0.037	0.177
	BL	0.156	0.104	0.687			MGB	0.085	0.091	0.236
	BGB	0.119	0.109	0.763			Multiple	0.170	0.158	0.490
	MP	0.136	0.120	0.176			Unknown	0.030	0.046	0.171
	ML	0.011	0.009	0.192		2009	BP	0.089	0.027	0.207
	MGB	0.034	0.054	0.512			BL	0.136	0.268	0.132
	Multiple	0.381	0.422	0.856			BGB	0.213	0.224	0.809
	Unknown	0.037	0.064	0.531			MP	0.199	0.181	0.635
2007	BP	0.128	0.133	0.437			ML	0.020	0.024	0.275
	BL	0.186	0.200	0.275			MGB	0.036	0.105	0.184
	BGB	0.075	0.085	0.634			Multiple	0.717	0.743	0.948
	MP	0.078	0.083	0.189			Unknown	0.178	0.166	0.531

A.2 Plausible relations - organ, age and year

Table A-4 The cancerous organs for patients corresponding to the age group and year combination

Age	**Year**	**P(data point | Organ =**			**Total**		**Age**	**Year**	**P(data point | Organ =**			**Total**
		Bile duct	Gallbladder	Liver					Bile duct	Gallbladder	Liver	
Under 46	2005	2	2	32	54		>60-66					
	2006	3	1	8	22			2008	11	4	25	68
	2007	6	2	30	54			2009	2	2	4	12
	2008	9	2	32	69		>66-71	2005	12	3	30	67
	2009	1	2	5	12			2006	6	3	10	24
46-54	2005	6	1	26	49			2007	5	4	16	49
	2006	4	0	6	22			2008	11	3	27	69
	2007	3	0	23	45			2009	1	2	10	18
	2008	8	1	27	75		>71-77	2005	8	5	15	44
	2009	0	0	9	15			2006	4	2	18	37
>54-60	2005	10	1	17	50			2007	10	3	32	66
	2006	2	1	17	28			2008	14	5	24	62
	2007	7	4	24	57			2009	4	1	13	22
	2008	4	4	15	55		77+	2005	5	5	18	44
	2009	1	1	9	12			2006	2	2	10	22
>60-66	2005	9	3	21	62			2007	15	2	27	66
	2006	4	1	14	25			2008	13	4	25	71
	2007	8	4	34	66			2009	0	2	5	13

Table A-5 The posterior probability of each data point for each learned parameter of H1: Organ is independent of Age and Year

Age	Year	P(data point | Organ =			Age	Year	P(data point | Organ =		
		Bile duct	Gallbladder	Liver			Bile duct	Gallbladder	Liver
Under 46	2005	0.013	0.228	0.007	>60-66				
	2006	0.238	0.365	0.141		2008	0.109	0.192	0.057
	2007	0.141	0.228	0.021		2009	0.283	0.111	0.187
	2008	0.135	0.164	0.081	>66-71	2005	0.081	0.210	0.092
	2009	0.322	0.111	0.226		2006	0.065	0.100	0.160
46-54	2005	0.160	0.184	0.044		2007	0.137	0.146	0.040
	2006	0.182	0.298	0.058		2008	0.112	0.205	0.077
	2007	0.074	0.088	0.067		2009	0.199	0.183	0.107
	2008	0.105	0.070	0.044	>71-77	2005	0.110	0.059	0.060
	2009	0.109	0.436	0.088		2006	0.176	0.272	0.103
>54-60	2005	0.068	0.178	0.050		2007	0.126	0.212	0.066
	2006	0.152	0.333	0.028		2008	0.023	0.124	0.080
	2007	0.147	0.173	0.103		2009	0.182	0.365	0.056
	2008	0.065	0.167	0.007	77+	2005	0.164	0.059	0.114
	2009	0.322	0.349	0.021		2006	0.224	0.220	0.163
>60-66	2005	0.137	0.219	0.036		2007	0.019	0.177	0.091
	2006	0.201	0.351	0.070		2008	0.069	0.194	0.040
	2007	0.134	0.190	0.039		2009	0.146	0.124	0.209

Table A-6 The posterior probability of each data point for each learned parameter of H2: Organ is dependent on both Age and Year

Age	Year	P(data point | Organ =			Age	Year	P(data point | Organ =		
		Bile duct	Gallbladder	Liver			Bile duct	Gallbladder	Liver
Under 46	2005	0.191	0.191	0.077	>60-66				
	2006	0.171	0.259	0.124		2008	0.092	0.141	0.070
	2007	0.119	0.191	0.077		2009	0.211	0.211	0.172
	2008	0.099	0.190	0.067	>66-71	2005	0.089	0.159	0.068
	2009	0.270	0.211	0.165		2006	0.131	0.169	0.116
46-54	2005	0.121	0.254	0.080		2007	0.130	0.143	0.085
	2006	0.152	0.510	0.134		2008	0.091	0.159	0.069
	2007	0.162	0.504	0.084		2009	0.263	0.203	0.134
	2008	0.104	0.251	0.067	>71-77	2005	0.108	0.131	0.089
	2009	0.515	0.515	0.148		2006	0.145	0.194	0.092
>54-60	2005	0.098	0.253	0.084		2007	0.095	0.160	0.069
	2006	0.196	0.257	0.108		2008	0.084	0.128	0.073

	2007	0.112	0.142	0.075			2009	0.152	0.259	0.122
	2008	0.142	0.142	0.085		77+	2005	0.131	0.131	0.085
	2009	0.270	0.270	0.185			2006	0.200	0.200	0.120
>60-66	2005	0.099	0.160	0.075			2007	0.082	0.191	0.070
	2006	0.151	0.258	0.113			2008	0.085	0.141	0.069
	2007	0.105	0.141	0.068			2009	0.517	0.210	0.160

Table A-7 The posterior probability of each data point for each learned parameter of H3: Organ is dependent of Age

Age	**Year**	**P(data point | Organ =**				**Age**	**Year**	**P(data point | Organ =**		
		Bile duct	Gallbladder	Liver				Bile duct	Gallbladder	Liver
Under 46	2005	0.067	0.237	0.052		>60-66				
	2006	0.202	0.357	0.072			2008	0.109	0.175	0.064
	2007	0.147	0.237	0.079			2009	0.284	0.138	0.193
	2008	0.098	0.198	0.068		>66-71	2005	0.099	0.155	0.073
	2009	0.364	0.091	0.184			2006	0.094	0.146	0.154
46-54	2005	0.143	0.305	0.055			2007	0.102	0.174	0.060
	2006	0.120	0.736	0.053			2008	0.113	0.149	0.082
	2007	0.153	0.552	0.075			2009	0.160	0.224	0.088
	2008	0.125	0.316	0.038		>71-77	2005	0.140	0.112	0.053
	2009	0.198	0.809	0.099			2006	0.109	0.238	0.106
>54-60	2005	0.047	0.166	0.073			2007	0.108	0.149	0.071
	2006	0.189	0.310	0.020			2008	0.071	0.156	0.068
	2007	0.139	0.165	0.092			2009	0.205	0.319	0.067
	2008	0.103	0.162	0.020		77+	2005	0.117	0.112	0.108
	2009	0.343	0.348	0.016			2006	0.172	0.255	0.138
>60-66	2005	0.126	0.186	0.047			2007	0.052	0.103	0.084
	2006	0.199	0.319	0.063			2008	0.098	0.156	0.070
	2007	0.112	0.176	0.035			2009	0.103	0.177	0.215

Table A-8 The posterior probability of each data point for each learned parameter of H4: Organ is dependent of Year

Age	Year	P(data point | Organ =				Age	Year	P(data point | Organ =		
		Bile duct	Gallbladder	Liver				Bile duct	Gallbladder	Liver
Under 46	2005	0.014	0.215	0.009		>60-66				
	2006	0.226	0.343	0.112			0.117	0.178	0.092	0.117
	2007	0.141	0.238	0.044			0.207	0.224	0.099	0.207
	2008	0.119	0.182	0.033		>66-71	0.085	0.195	0.086	0.085
	2009	0.362	0.224	0.165			0.074	0.118	0.140	0.074
46-54	2005	0.149	0.178	0.044			0.140	0.125	0.022	0.140
	2006	0.179	0.275	0.041			0.118	0.205	0.087	0.118
	2007	0.083	0.116	0.091			0.305	0.262	0.167	0.305
	2008	0.081	0.092	0.086		>71-77	0.113	0.067	0.062	0.113
	2009	0.249	0.215	0.168			0.162	0.251	0.113	0.162
>54-60	2005	0.075	0.172	0.053			0.117	0.209	0.084	0.117
	2006	0.145	0.305	0.051			0.039	0.110	0.094	0.039
	2007	0.143	0.150	0.084			0.097	0.233	0.134	0.097
	2008	0.048	0.151	0.037		77+	0.153	0.067	0.110	0.153
	2009	0.362	0.356	0.079			0.212	0.227	0.158	0.212
>60-66	2005	0.130	0.204	0.039			0.019	0.200	0.067	0.019
	2006	0.192	0.326	0.097			0.088	0.181	0.086	0.088
	2007	0.131	0.170	0.065			0.296	0.236	0.130	0.296

A.3 Plausible relations - type, age and year

Table A-9 The cancer severity stages for patients corresponding to the age group and year combination

Age	Year	P(data point | Type =		Total		Age	Year	P(data point | Type =		Total
		Benign	Malignant					Benign	Malignant	
Under 46	2005	35	19	54		>60-66				
	2006	13	9	22			2008	29	36	68
	2007	33	21	54			2009	4	8	12
	2008	47	20	69		>66-71	2005	15	51	67
	2009	9	3	12			2006	8	15	24
46-54	2005	20	29	49			2007	15	34	49
	2006	15	7	22			2008	20	47	69
	2007	18	27	45			2009	6	11	18
	2008	41	29	75		>71-77	2005	17	27	44
	2009	11	4	15			2006	10	27	37

>54-60	2005	16	33	50			2007	10	56	66
	2006	11	17	28			2008	12	45	62
	2007	21	34	57			2009	5	15	22
	2008	19	32	55		77+	2005	18	25	44
	2009	5	6	12			2006	5	16	22
>60-66	2005	18	42	62			2007	14	51	66
	2006	6	19	25			2008	23	41	71
	2007	14	51	66			2009	4	9	13

Table A-10 The posterior probability of each data point for each learned parameter of H1: Type is independent Age and Year

Age	**Year**	**P(data point | Type =**			**Age**	**Year**	**P(data point | Type =**	
		Benign	Malignant				Benign	Malignant
Under 46	2005	3.42E-05	2.09E-04		>60-66			
	2006	0.020	0.034			2008	0.063	0.048
	2007	2.51E-04	0.001			2009	0.228	0.212
	2008	3.83E-07	7.75E-05		>66-71	2005	0.005	0.003
	2009	0.008	0.013			2006	0.157	0.160
46-54	2005	0.099	0.112			2007	0.077	0.049
	2006	0.003	0.005			2008	0.039	0.040
	2007	0.110	0.118			2009	0.184	0.188
	2008	0.001	1.40E-04		>71-77	2005	0.118	0.119
	2009	0.004	0.008			2006	0.063	0.038
>54-60	2005	0.089	0.081			2007	1.29E-04	1.34E-05
	2006	0.147	0.151			2008	0.002	0.014
	2007	0.107	0.105			2009	0.070	0.131
	2008	0.102	0.102		77+	2005	0.105	0.108
	2009	0.216	0.176			2006	0.070	0.087
>60-66	2005	0.045	0.050			2007	0.003	0.002
	2006	0.069	0.045			2008	0.070	0.086
	2007	0.003	0.002			2009	0.207	0.184

Table A-11 The posterior probability of each data point for each learned parameter of H2: Type is dependent on Age and Year

Age	Year	P(data point | Type =			Age	Year	P(data point | Type =	
		Benign	Malignant				Benign	Malignant
Under 46	2005	0.078	0.079		>60-66			
	2006	0.122	0.121			2008	0.068	0.068
	2007	0.078	0.078			2009	0.172	0.172
	2008	0.072	0.074		>66-71	2005	0.081	0.080
	2009	0.185	0.186			2006	0.122	0.118
46-54	2005	0.080	0.081			2007	0.087	0.087
	2006	0.128	0.128			2008	0.074	0.072
	2007	0.085	0.085			2009	0.140	0.136
	2008	0.065	0.066		>71-77	2005	0.086	0.087
	2009	0.163	0.163			2006	0.103	0.103
>54-60	2005	0.084	0.083			2007	0.095	0.095
	2006	0.109	0.108			2008	0.089	0.079
	2007	0.077	0.076			2009	0.142	0.128
	2008	0.079	0.076		77+	2005	0.085	0.085
	2009	0.165	0.163			2006	0.142	0.134
>60-66	2005	0.078	0.076			2007	0.084	0.082
	2006	0.130	0.130			2008	0.071	0.066
	2007	0.084	0.082			2009	0.169	0.168

Table A-12 The posterior probability of each data point for each learned parameter of H3: Type is dependent on Age

Age	Year	P(data point | Type =			Age	Year	P(data point | Type =	
		Benign	Malignant				Benign	Malignant
Under 46	2005	0.100	0.099		>60-66			
	2006	0.140	0.133			2008	0.017	0.010
	2007	0.086	0.080			2009	0.226	0.231
	2008	0.078	0.066		>66-71	2005	0.064	0.057
	2009	0.191	0.200			2006	0.141	0.124
46-54	2005	0.046	0.030			2007	0.105	0.111
	2006	0.051	0.070			2008	0.090	0.087
	2007	0.045	0.029			2009	0.170	0.140

	2008	0.067	0.040		>71-77	2005	0.015	0.032
	2009	0.051	0.068			2006	0.121	0.134
>54-60	2005	0.095	0.080			2007	0.039	0.017
	2006	0.133	0.142			2008	0.088	0.098
	2007	0.094	0.093			2009	0.189	0.149
	2008	0.098	0.091		77+	2005	0.040	0.060
	2009	0.203	0.171			2006	0.148	0.141
>60-66	2005	0.094	0.094			2007	0.040	0.020
	2006	0.136	0.111			2008	0.079	0.041
	2007	0.034	0.027			2009	0.226	0.218

Table A-13 The posterior probability of each data point for each learned parameter of H3: Type is dependent on Year

Age	**Year**	**P(data point | Type =**			**Age**	**Year**	**P(data point | Type =**	
		Benign	Malignant				Benign	Malignant
Under 46	2005	1.01E-04	3.62E-04		>60-66			
	2006	0.027	0.033			2008	0.086	0.090
	2007	2.07E-05	3.36E-05			2009	0.187	0.156
	2008	3.30E-05	7.94E-05		>66-71	2005	0.005	0.006
	2009	0.023	0.039			2006	0.145	0.154
46-54	2005	0.098	0.104			2007	0.115	0.113
	2006	0.004	0.006			2008	0.017	0.006
	2007	0.056	0.065			2009	0.139	0.148
	2008	0.007	0.006		>71-77	2005	0.114	0.115
	2009	0.016	0.029			2006	0.059	0.051
>54-60	2005	0.083	0.088			2007	0.003	0.002
	2006	0.139	0.141			2008	0.000	0.002
	2007	0.070	0.047			2009	0.038	0.075
	2008	0.071	0.081		77+	2005	0.104	0.098
	2009	0.215	0.206			2006	0.067	0.098
>60-66	2005	0.043	0.060			2007	0.028	0.032
	2006	0.066	0.057			2008	0.038	0.069
	2007	0.028	0.032			2009	0.158	0.124

Appendix B

Evaluating the strength of relations

B.1 Impact of age on cancer types

Table B-1 Bayes Factors (BFs) to analyse the impact of Age on Type

Compared to	The probability is high for	BFs		
		Benign	Malignant	Unknown
Under 46	46-54	0.004	59.184	1.276
	>54-60	3.54E-05	471.524	2.984
	>60-66	2.43E-05	5040.888	1.338
	>66-71	4.09E-05	912.245	1.119
	>71-77	9.13E-06	1508.999	1.835
	77+	5.32E-05	836.898	4.223
46-54	>54-60	0.004	63.898	2.177
	>60-66	3.09E-04	520.707	0.847
	>66-71	0.002	859.361	0.716
	>71-77	2.25E-04	1473.737	1.252
	77+	3.24E-04	380.113	3.193
>54-60	>60-66	0.140	7.984	0.341
	>66-71	0.055	21.063	0.272
	>71-77	0.005	123.453	0.507
	77+	0.107	4.652	1.208
>60-66	>66-71	0.394	2.132	0.688
	>71-77	0.047	9.449	1.229
	77+	0.654	0.603	3.280
>66-71	>71-77	0.140	3.450	1.625
	77+	1.467	0.263	4.536
71-77	77+	9.885	0.053	2.174

B.2 Impact of year on type, organ and treatment

Table B-2 BFs to analyse the impact of Year on Type

Compared to	The probability is high for	BFs		
		Benign	Malignant	Unknown
2005	2006	0.928	0.782	1.023
	2007	0.028	19.975	0.601
	2008	3.725	0.026	40.241
	2009	3.23	0.101	6.066
2006	2007	0.06	11.712	0.522
	2008	2.558	0.06	20.099
	2009	2.623	0.138	4.479
2007	2008	150.196	0.003	53.275
	2009	33.753	0.01	7.748
2008	2009	1.263	0.829	0.452

Table B-3 BFs to analyse the impact of Year on Organ

Compared to	The probability is high for	BFs			
		Pancreas	Bile duct	Gallbladder	Liver
2005	2006	0.31	0.81	1.04	2.21
	2007	0.4	0.6	0.43	3.33
	2008	9.08	1.47	0.51	0.07
	2009	0.05	0.08	11.28	13.61
2006	2007	1.34	0.7	0.4	0.98
	2008	20.04	1.45	0.46	0.03
	2009	0.19	0.11	7.31	4.29
2007	2008	22.41	2.33	0.96	0.01
	2009	0.09	0.11	20.58	5.01
2008	2009	0.01	0.05	19.16	92.95

Table B-4 BFs to analyse the impact of Year on Treatment

Compared to	The probability is high for	BFs			
		Chemotherapy	Combination	Palliative	Surgery
2005	2006	0.065	0.095	11.610	3.375
	2007	7.168	9.03E-04	5.941	34.760
	2008	0.524	6.70E-04	0.300	33.856
	2009	2.626	3.47E-04	1.660	1.204
2006	2007	97.639	0.045	0.362	4.118
	2008	7.944	0.015	0.022	4.059
	2009	21.477	0.002	0.257	0.424
2007	2008	0.044	0.313	0.033	0.825
	2009	0.638	0.014	0.489	0.094
2008	2009	3.698	0.032	3.657	0.094

B.3 Impact of diagnosis on treatment and MDT meetings

Table B-5 BFs to analyse the impact of Diagnosis on Surgery

Compared to	The probability is high for	BFs
		Surgery
BP	BL	0.047
	BGB	1690.196
	MP	3395.095
	ML	98.949
	MGB	4865.029
	Multiple	1045.525
	Unknown	0.001
BL	BGB	24887.508
	MP	16250.131
	ML	1351.473
	MGB	67662.794
	Multiple	8252.610
	Unknown	0.030
BGB	MP	1.072
	ML	0.038
	MGB	11.355
	Multiple	3.149
	Unknown	1.05E-06
MP	ML	0.017
	MGB	14.098

	Multiple	3.183
	Unknown	6.54E-06
ML	MGB	416.795
	Multiple	38.551
	Unknown	1.6E-05
MGB	Multiple	0.555
	Unknown	5.94E-06
Multiple	Unknown	1.23E-05

Table B-6 BFs to analyse the impact of Diagnosis on Number of meetings

Compared to	**The probability is high for**	**BFs**			
		1	2	3	4 or more
BP	BL	4.342	0.140	0.758	2.981
	BGB	0.699	1.103	0.671	1.633
	MP	2.422	0.306	1.018	1.115
	ML	0.134	2.795	2.365	2.730
	MGB	0.094	5.687	1.732	0.878
	Multiple	1.001	0.384	1.979	2.905
	Unknown	2.294	0.256	1.171	1.756
BL	BGB	0.177	6.315	0.781	0.522
	MP	0.403	2.811	1.271	0.274
	ML	0.018	31.481	3.016	0.613
	MGB	0.015	49.561	2.111	0.238
	Multiple	0.350	1.421	2.382	1.108
	Unknown	0.462	1.909	1.471	0.480
BGB	MP	2.821	0.276	1.447	0.549
	ML	0.252	1.842	3.199	1.369
	MGB	0.160	3.657	2.270	0.481
	Multiple	1.178	0.343	2.514	1.816
	Unknown	2.657	0.234	1.644	0.974
MP	ML	0.012	20.590	2.501	3.004
	MGB	0.013	35.741	1.705	0.684
	Multiple	0.562	0.712	1.991	2.996
	Unknown	0.979	0.628	1.058	1.684
ML	MGB	0.401	2.689	0.755	0.252
	Multiple	2.966	0.172	1.147	1.382
	Unknown	32.239	0.041	0.442	0.578
MGB	Multiple	4.214	0.096	1.233	3.461
	Unknown	37.625	0.023	0.574	2.078
Multiple	Unknown	1.493	0.996	0.461	0.434